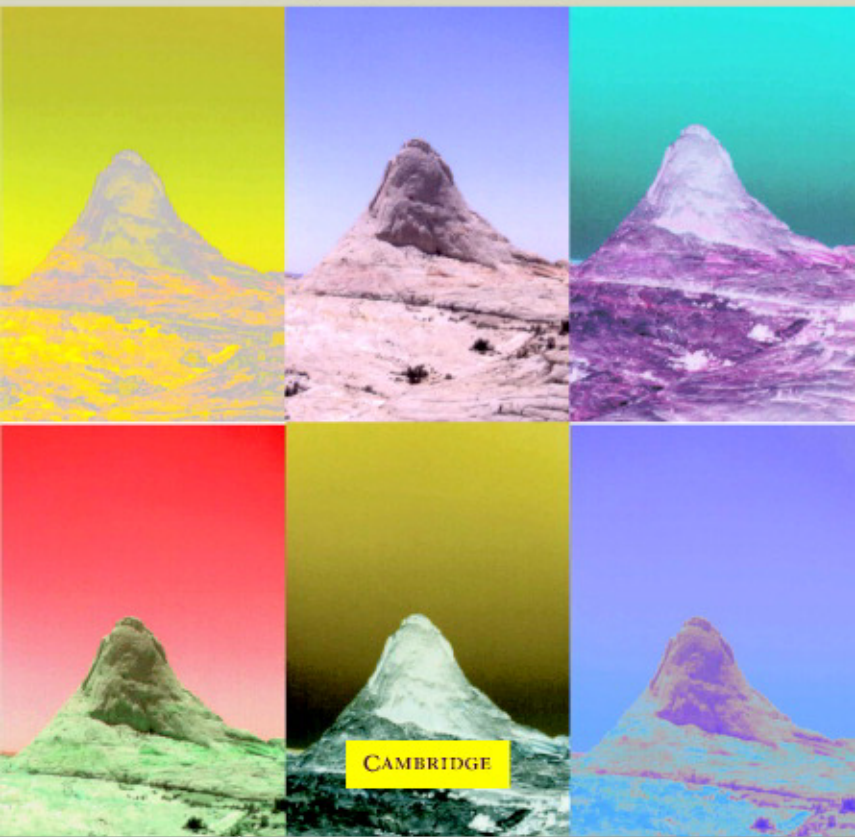


ELEMENTARY PROBABILITY

SECOND EDITION

David Stirzaker



CAMBRIDGE

more information - www.cambridge.org/9780521833448

This page intentionally left blank

Elementary Probability

2nd Edition

Now available in a fully revised and updated new edition, this well-established textbook provides a straightforward introduction to the theory of probability. The presentation is entertaining without any sacrifice of rigour; important notions are covered with the clarity that the subject demands.

Topics covered include conditional probability, independence, discrete and continuous random variables, basic combinatorics, generating functions and limit theorems, and an introduction to Markov chains. This edition includes an elementary approach to martingales and the theory of Brownian motion, which supply the cornerstones for many topics in modern financial mathematics such as option and derivative pricing. The text is accessible to undergraduate students, and provides numerous worked examples and exercises to help build the important skills necessary for problem solving.

‘[T]he author succeeds in combining the utmost succinctness with clarity and genuine readability. . . . This textbook can be recommended unreservedly.’

Internationale Mathematische Nachrichten

‘[T]his book is a superb resource of theory and application, which should be on every lecturer’s shelves, and those of many students. You may never need to buy another book on probability.’

Keith Hirst, *The Mathematical Gazette*

‘Excellent! A vast number of well-chosen worked examples and exercises guide the reader through the basic theory of probability at the elementary level . . . an excellent text which I am sure will give a lot of pleasure to students and teachers alike.’

International Statistics Institute

‘[W]ould make a fine addition to an undergraduate library. A student with a solid background in calculus, linear algebra, and set theory will find many useful tools of elementary probability here.’

Phil Gilbert, *The Mathematics Teacher*

‘Stirzaker does an excellent job of developing problem-solving skills in an introductory probability text. Numerous examples and practice exercises are provided that only serve to enhance a student’s problem-solving abilities. . . . Highly recommended.’

D.J. Gougeon, *Choice*

‘The book would make an excellent text for the properly prepared class, a solid instructor’s reference for both probability applications and problems, as well as a fine work for purposes of self-study.’

J. Philip Smith, *School Science and Mathematics*

Elementary Probability

2nd Edition

by

DAVID STIRZAKER

*Mathematical Institute and St. John's College,
University of Oxford*



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS

Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press

The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521833448

© David Stirzaker 2003

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

First published in print format 2003

ISBN-13 978-0-521-07574-2 eBook (Adobe Reader)

ISBN-10 0-521-07574-X eBook (Adobe Reader)

ISBN-13 978-0-521-83344-8 hardback

ISBN-10 0-521-83344-2 hardback

ISBN-13 978-0-521-53428-4 paperback

ISBN-10 0-521-53428-3 paperback

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

Contents

Preface to the Second Edition	<i>page xi</i>
0 Introduction	1
0.1 Chance	1
0.2 Models	3
0.3 Symmetry	5
0.4 The Long Run	7
0.5 Pay-Offs	8
0.6 Introspection	9
0.7 FAQs	10
0.8 History	14
Appendix: Review of Elementary Mathematical Prerequisites	15
1 Probability	24
1.1 Notation and Experiments	24
1.2 Events	26
1.3 The Addition Rules for Probability	32
1.4 Properties of Probability	34
1.5 Sequences of Events	36
1.6 Remarks	37
1.7 Review and Checklist for Chapter 1	38
Worked examples and exercises	40
1.8 Example: Dice	40
1.9 Example: Urn	41
1.10 Example: Cups and Saucers	42
1.11 Example: Sixes	43
1.12 Example: Family Planning	44
1.13 Example: Craps	45
1.14 Example: Murphy's Law	46
Problems	47
2 Conditional Probability and Independence	51
2.1 Conditional Probability	51
2.2 Independence	57
2.3 Recurrence and Difference Equations	60
2.4 Remarks	62

2.5	Review and Checklist for Chapter 2	64
	Worked examples and exercises	65
2.6	Example: Sudden Death	65
2.7	Example: Polya's Urn	66
2.8	Example: Complacency	67
2.9	Example: Dogfight	68
2.10	Example: Smears	69
2.11	Example: Gambler's Ruin	70
2.12	Example: Accidents and Insurance	72
2.13	Example: Protocols	73
2.14	Example: Eddington's Controversy	75
	Problems	76
3	Counting	83
3.1	First Principles	83
3.2	Permutations: Ordered Selection	84
3.3	Combinations: Unordered Selection	86
3.4	Inclusion–Exclusion	87
3.5	Recurrence Relations	88
3.6	Generating Functions	90
3.7	Techniques	93
3.8	Review and Checklist for Chapter 3	95
	Worked examples and exercises	97
3.9	Example: Railway Trains	97
3.10	Example: Genoese Lottery	98
3.11	Example: Ringing Birds	99
3.12	Example: Lottery	101
3.13	Example: The Ménages Problem	101
3.14	Example: Identity	102
3.15	Example: Runs	103
3.16	Example: Fish	105
3.17	Example: Colouring	106
3.18	Example: Matching (Rencontres)	107
	Problems	108
4	Random Variables: Distribution and Expectation	114
4.1	Random Variables	114
4.2	Distributions	115
4.3	Expectation	120
4.4	Conditional Distributions	127
4.5	Sequences of Distributions	130
4.6	Inequalities	131
4.7	Review and Checklist for Chapter 4	134
	Worked examples and exercises	137
4.8	Example: Royal Oak Lottery	137
4.9	Example: Misprints	138
4.10	Example: Dog Bites: Poisson Distribution	139

4.11	Example: Guesswork	141
4.12	Example: Gamblers Ruined Again	142
4.13	Example: Postmen	143
4.14	Example: Acme Gadgets	144
4.15	Example: Roulette and the Martingale	145
4.16	Example: Searching	146
4.17	Example: Duelling	147
4.18	Binomial Distribution: The Long Run	149
4.19	Example: Uncertainty and Entropy	150
	Problems	151
5	Random Vectors: Independence and Dependence	158
5.1	Joint Distributions	158
5.2	Independence	162
5.3	Expectation	165
5.4	Sums and Products of Random Variables: Inequalities	172
5.5	Dependence: Conditional Expectation	177
5.6	Simple Random Walk	183
5.7	Martingales	190
5.8	The Law of Averages	196
5.9	Convergence	199
5.10	Review and Checklist for Chapter 5	203
	Worked examples and exercises	206
5.11	Example: Golf	206
5.12	Example: Joint Lives	208
5.13	Example: Tournament	209
5.14	Example: Congregations	210
5.15	Example: Propagation	211
5.16	Example: Information and Entropy	212
5.17	Example: Cooperation	214
5.18	Example: Strange But True	215
5.19	Example: Capture–Recapture	216
5.20	Example: Visits of a Random Walk	218
5.21	Example: Ordering	219
5.22	Example: More Martingales	220
5.23	Example: Simple Random Walk Martingales	221
5.24	Example: You Can’t Beat the Odds	222
5.25	Example: Matching Martingales	223
5.26	Example: Three-Handed Gambler’s Ruin	224
	Problems	226
6	Generating Functions and Their Applications	232
6.1	Introduction	232
6.2	Moments and the Probability Generating Function	236
6.3	Sums of Independent Random Variables	239
6.4	Moment Generating Functions	245
6.5	Joint Generating Functions	247

6.6	Sequences	251
6.7	Regeneration	254
6.8	Random Walks	259
6.9	Review and Checklist for Chapter 6	263
	Appendix: Calculus	265
	Worked examples and exercises	268
6.10	Example: Gambler's Ruin and First Passages	268
6.11	Example: "Fair" Pairs of Dice	269
6.12	Example: Branching Process	271
6.13	Example: Geometric Branching	272
6.14	Example: Waring's Theorem: Occupancy Problems	274
6.15	Example: Bernoulli Patterns and Runs	275
6.16	Example: Waiting for Unusual Light Bulbs	277
6.17	Example: Martingales for Branching	278
6.18	Example: Wald's Identity	279
6.19	Example: Total Population in Branching Problems	281
7	Continuous Random Variables	287
7.1	Density and Distribution	287
7.2	Functions of Random Variables	297
7.3	Simulation of Random Variables	301
7.4	Expectation	302
7.5	Moment Generating Functions	306
7.6	Conditional Distributions	310
7.7	Ageing and Survival	312
7.8	Stochastic Ordering	314
7.9	Random Points	315
7.10	Review and Checklist for Chapter 7	318
	Worked examples and exercises	321
7.11	Example: Using a Uniform Random Variable	321
7.12	Example: Normal Distribution	323
7.13	Example: Bertrand's Paradox	324
7.14	Example: Stock Control	326
7.15	Example: Obtaining Your Visa	327
7.16	Example: Pirates	329
7.17	Example: Failure Rates	330
7.18	Example: Triangles	330
7.19	Example: Stirling's Formula Problems	332
8	Jointly Continuous Random Variables	337
8.1	Joint Density and Distribution	337
8.2	Change of Variables	342
8.3	Independence	344
8.4	Sums, Products, and Quotients	348
8.5	Expectation	351
8.6	Conditional Density and Expectation	355

8.7	Transformations: Order Statistics	361
8.8	The Poisson Process: Martingales	364
8.9	Two Limit Theorems	368
8.10	Review and Checklist for Chapter 8	371
	Worked examples and exercises	375
8.11	Example: Bivariate Normal Density	375
8.12	Example: Partitions	376
8.13	Example: Buffon's Needle	377
8.14	Example: Targets	379
8.15	Example: Gamma Densities	380
8.16	Example: Simulation – The Rejection Method	381
8.17	Example: The Inspection Paradox	382
8.18	Example: von Neumann's Exponential Variable	383
8.19	Example: Maximum from Minima	385
8.20	Example: Binormal and Trinormal	387
8.21	Example: Central Limit Theorem	388
8.22	Example: Poisson Martingales	389
8.23	Example: Uniform on the Unit Cube	390
8.24	Example: Characteristic Functions	390
	Problems	391
9	Markov Chains	396
9.1	The Markov Property	396
9.2	Transition Probabilities	400
9.3	First Passage Times	406
9.4	Stationary Distributions	412
9.5	The Long Run	418
9.6	Markov Chains with Continuous Parameter	425
9.7	Forward Equations: Poisson and Birth Processes	428
9.8	Forward Equations: Equilibrium	431
9.9	The Wiener Process and Diffusions	436
9.10	Review and Checklist for Chapter 9	449
	Worked examples and exercises	451
9.11	Example: Crossing a Cube	451
9.12	Example: Reversible Chains	453
9.13	Example: Diffusion Models	454
9.14	Example: The Renewal Chains	456
9.15	Example: Persistence	457
9.16	Example: First Passages and Bernoulli Patterns	459
9.17	Example: Poisson Processes	461
9.18	Example: Decay	462
9.19	Example: Disasters	463
9.20	Example: The General Birth Process	465
9.21	Example: The Birth–Death Process	466
9.22	Example: Wiener Process with Drift	468
9.23	Example: Markov Chain Martingales	469
9.24	Example: Wiener Process Exiting a Strip	470

9.25	Example: Arcsine Law for Zeros	471
9.26	Example: Option Pricing: Black–Scholes Formula	472
	Problems	473
	Appendix: Solutions and Hints for Selected Exercises and Problems	478
	Further Reading	514
	Index of Notation	515
	Index	517

Preface to the Second Edition

The calculus of probabilities, in an appropriate form, should interest equally the mathematician, the experimentalist, and the statesman. . . . It is under its influence that lotteries and other disgraceful traps cunningly laid for greed and ignorance have finally disappeared.

Francois Arago, Eulogy on Laplace, 1827

Lastly, one of the principal uses to which this Doctrine of Chances may be applied, is the discovering of some truths, which cannot fail of pleasing the mind, by their generality and simplicity; the admirable connexion of its consequences will increase the pleasure of the discovery; and the seeming paradoxes wherewith it abounds, will afford very great matter of surprize and entertainment to the inquisitive.

Abraham de Moivre, The Doctrine of Chances, 1756

This book provides an introduction to elementary probability and some of its simple applications. In particular, a principal purpose of the book is to help the student to solve problems. Probability is now being taught to an ever wider audience, not all of whom can be assumed to have a high level of problem-solving skills and mathematical background. It is also characteristic of probability that, even at an elementary level, few problems are entirely routine. Successful problem solving requires flexibility and imagination on the part of the student. Commonly, these skills are developed by observation of examples and practice at exercises, both of which this text aims to supply.

With these targets in mind, in each chapter of the book, the theoretical exposition is accompanied by a large number of examples and is followed by worked examples incorporating a cluster of exercises. The examples and exercises have been chosen to illustrate the subject, to help the student solve the kind of problems typical of examinations, and for their entertainment value. (Besides its practical importance, probability is without doubt one of the most entertaining branches of mathematics.) Each chapter concludes with problems: solutions to many of these appear in an appendix, together with the solutions to most of the exercises.

The ordering and numbering of material in this second edition has for the most part been preserved from the first. However, numerous alterations and additions have been included to make the basic material more accessible and the book more useful for self-study. In

particular, there is an entirely new introductory chapter that discusses our informal and intuitive ideas about probability, and explains how (and why) these should be incorporated into the theoretical framework of the rest of the book. Also, all later chapters now include a section entitled, “Review and checklist,” to aid the reader in navigation around the subject, especially new ideas and notation.

Furthermore, a new section of the book provides a first introduction to the elementary properties of martingales, which have come to occupy a central position in modern probability. Another new section provides an elementary introduction to Brownian motion, diffusion, and the Wiener process, which has underpinned much classical financial mathematics, such as the Black–Scholes formula for pricing options. Optional stopping and its applications are introduced in the context of these important stochastic models, together with several associated new worked examples and exercises.

The basic structure of the book remains unchanged; there are three main parts, each comprising three chapters.

The first part introduces the basic ideas of probability, conditional probability, and independence. It is assumed that the reader has some knowledge of elementary set theory. (We adopt the now conventional formal definition of probability. This is not because of high principles, but merely because the alternative intuitive approach seems to lead more students into errors.) The second part introduces discrete random variables, probability mass functions, and expectation. It is assumed that the reader can do simple things with functions and series. The third part considers continuous random variables, and for this a knowledge of the simpler techniques of calculus is desirable.

In addition, there are chapters on combinatorial methods in probability, the use of probability (and other) generating functions, and the basic theory of Markov processes in discrete and continuous time. These sections can be omitted at a first reading, if so desired.

In general, the material is presented in a conventional order, which roughly corresponds to increasing levels of knowledge and dexterity on the part of the reader. Those who start with a sufficient level of basic skills have more freedom to choose the order in which they read the book. For example, you may want to read Chapters 4 and 7 together (and then Chapters 5 and 8 together), regarding discrete and continuous random variables as two varieties of the same species (which they are). Also, much of Chapter 9 could be read immediately after Chapter 5, if you prefer.

In particular, the book is structured so that the first two parts are suitable to accompany the probability component of a typical course in discrete mathematics; a knowledge of calculus is not assumed until the final part of the book. This layout entails some repetition of similar ideas in different contexts, and this should help to reinforce the reader’s knowledge of the less elementary concepts and techniques.

The ends of examples, proofs, and definitions are indicated by the symbols ●, ■, and ▲, respectively.

Finally, you should note that the book contains a random number of errors. I entreat readers to inform me of all those they find.

0

Introduction

A life which included no improbable events would be the real statistical improbability.

Poul Anderson

It is plain that any scientist is trying to correlate the incoherent body of facts confronting him with some definite and orderly scheme of abstract relations, the kind of scheme which he can borrow only from mathematics.

G.H. Hardy

This chapter introduces the basic concepts of probability in an informal way. We discuss our everyday experience of chance, and explain why we need a theory and how we start to construct one. Mathematical probability is motivated by our intuitive ideas about likelihood as a proportion in many practical instances. We discuss some of the more common questions and problems in probability, and conclude with a brief account of the history of the subject.

0.1 Chance

My only solution for the problem of habitual accidents is to stay in bed all day. Even then, there is always the chance that you will fall out.

Robert Benchley

It is not certain that everything is uncertain.

Blaise Pascal

You can be reasonably confident that the sun will rise tomorrow, but what it will be shining on is a good deal more problematical. In fact, the one thing we can be certain of is that uncertainty and randomness are unavoidable aspects of our experience.

At a personal level, minor ailments and diseases appear unpredictably and are resolved not much more predictably. Your income and spending are subject to erratic strokes of good or bad fortune. Your genetic makeup is a random selection from those of your parents. The weather is notoriously fickle in many areas of the globe. You may decide to play

cards, invest in shares, bet on horses, buy lottery tickets, or engage in one or several other forms of gambling on events that are necessarily uncertain (otherwise, gambling could not occur).

At a different level, society has to organize itself in the context of similar sorts of uncertainty. Engineers have to build structures to withstand stressful events of unknown magnitude and frequency. Computing and communication systems need to be designed to cope with uncertain and fluctuating demands and breakdowns. Any system should be designed to have a small chance of failing and a high chance of performing as it was intended. Financial markets of any kind should function so as to share out risks in an efficient and transparent way, for example, when you insure your car or house, buy an annuity, or mortgage your house.

This uncertainty is not confined to the future and events that have yet to occur; much effort is expended by scientists (and by lawyers, curiously) who seek to resolve our doubt about things that have already occurred. Of course, our ignorance of the past is perhaps not quite as pressing as our ignorance of the future because of the direction in which time's arrow seems to be running. (But the arguments about the past are, paradoxically, somewhat more bad tempered as a rule.) In addition, and maybe most annoyingly, we are not certain about events occurring right now, even among those within our direct observation. At a serious level, you can see the human genome expressing itself in everyone you know, but the mechanisms remain largely a mystery. The task of unravelling this genetic conundrum will require a great deal of probability theory and statistical analysis. At a more trivial level, illusionists (and politicians) make a handsome living from our difficulties in being certain about our own personal experience (and prospects).

It follows that everyone must have some internal concept of chance to live in the real world, although such ideas may be implicit or even unacknowledged.

These concepts of chance have long been incorporated into many cultures in mythological or religious form. The casting of lots (sortilege) to make choices at random is widespread; we are all familiar with "the short straw" and the "lucky number." The Romans, for example, had gods of chance named Fortuna and Fors, and even today we have Lady Luck. Note that if you ransack the archives of the literary response to this state of affairs, one finds it to be extremely negative:

- "Fortune, that favours fools." *Ben Jonson*
- "For herein Fortune shows herself more kind than is her custom." *William Shakespeare, Merchant of Venice*
- "Ill fortune seldom comes alone." *John Dryden*
- "[T]he story of my life... wherein I spake of most disastrous chances." *William Shakespeare, Othello*
- Probability is the bane of the age. *Anthony Powell, Casanova's Chinese Restaurant*

This list of reproaches to Fortune could be extended almost indefinitely; in fact, you may have expressed similar sentiments yourself (although perhaps less poetically).

Nevertheless, it is a curious feature of human nature that, despite our oft-stated deprecation of this randomness, many people seek out extra uncertainty. They enter lotteries, bet on horses, and free-climb on rockfaces of dubious stability. A huge part of the entertainment industry is geared toward supplying surprise and uncertainty. This simultaneous desire to

be safe and yet at risk is an interesting trait that seems difficult to explain; fortunately, however, that is not our problem here.

Instead, our task is to find a way of describing and analysing the concepts of chance and uncertainty that we intuitively see are common to the otherwise remarkably diverse examples mentioned above.

0.2 Models

... and blessed are those
 whose blood and judgement are so well comingled
 that they are not a pipe for Fortune's finger
 to sound what stop she please.

W. Shakespeare, Hamlet

In the preceding section, we concluded that large parts of our experience are unpredictable and uncertain. To demonstrate the effect of chance in our lives, we gave a long list of examples, and we could have made it a great deal longer were it not for lack of space and fear of boring the reader. However, to say that most things are unpredictable is to paint too negative a picture. In fact, many things are certain (death and taxes, notoriously) and even uncertain things are susceptible to judgment and insight. We learn that, in Monopoly, it is good to own the orange set of properties; we know that casinos invariably make profits; we believe that it does not really matter whether you call heads or tails when a coin is flipped energetically enough; we learn not to be on top of the mountain during a thunderstorm; and so on.

In fact, we often go further than these rough judgments and compare probabilities. Most people would agree that in roulette, black is more likely than green (the zeros); a bookmaker is more likely to show a profit than a loss on a book; the chance of a thunderstorm is greater later in the day; and so on. This is another list that could be extended indefinitely, but the point is that *because* probabilities are often comparable in this way it is natural to represent them on a numerical scale. After all, such comparisons were the principal reason for the development of numbers in the first place. It will later be shown that this numerical scale should run from 0 to 1, but we first make some general remarks.

It seems that we do share a common concept of chance because we can discuss it and make agreed statements and judgments such as those above. We therefore naturally seek to abstract these essential common features, rather than discuss an endless list of examples from first principles. This type of simple (or at least, simplified) description of a system or concept is often called a model. Agreeing that probability is a number is the first step on our path to constructing our model.

Most, perhaps all, of science conforms to this pattern; astronomy was originally developed to describe the visible movements of planets and stars; Newton's and Einstein's theories of space, time, and motion were developed to describe our perceptions of moving bodies with their mass, energy, and motion; Maxwell's equations codify the properties of electromagnetism; and so on. The first advantage of such models is their concise description of otherwise incomprehensibly complicated systems.

The second, and arguably principal, advantage and purpose of having such a model is that (if it is well chosen) it provides not only a description of the system, but also predictions about how it will behave in the future. It may also predict how it would behave in different circumstances, or shed light on its (unobserved) past behaviour.

Astronomy is one example we have mentioned; for another, consider the weather. Without a model for forecasting, your only recourse is to recall the various ways in which weather developed on the previous occasions when the situation seemed to resemble the current one. There will almost certainly be no perfect match, and identifying a “good fit” will be exceedingly time consuming or impossible.

Returning to chance and probability, we note that a primitive model for chance, used by many cultures, represents it as a supernatural entity, or god. We mentioned this in the previous section, and this procedure is, from one point of view, a perfectly reasonable and consistent model for chance. It explains the data, with no contradictions. Unfortunately, it is useless for practical purposes, such as prediction and judgment, because it is necessary that the mind of the god in question should be unpredictable and capricious, and that mind of Fortune (or whatever) is closed to us. Efforts to discover Fortune’s inclination by propitiation of various kinds (sacrifice and wheedling) have met with outcomes that can at best be described as equivocal. The Romans made use of more complicated and various techniques, such as examining the behaviour of birds (augury) or casting lots (sortilege). Related modern techniques use tea leaves and astrology, but there is no evidence to suggest that any of these methods rate better than utterly useless.

Fortunately, experience over the past millennium has shown that we can do much better by using a mathematical model. This has many advantages; we mention only a few. First, a useful model must be simpler than reality; otherwise, it would be no easier to analyse than the real-life problem. Mathematical models have this stripped-down quality in abundance.

Second, mathematical models are abstract and are therefore quite unconstrained in their applications. When we define the probabilities of events in Chapter 1, and the rules that govern them, our conclusions will apply to all events of whatever kind (e.g., insurance claims, computer algorithms, crop failures, scientific experiments, games of chance; think of some more yourself).

Third, the great majority of practical problems about chance deal with questions that either are intrinsically numerical or can readily be rephrased in numerical terms. The use of a mathematical model becomes almost inescapable.

Fourth, if you succeed in constructing a model in mathematical form, then all the power of mathematics developed over several thousand years is instantly available to help you use it. Newton, Gauss, and Laplace become your (unpaid) assistants, and aides like these are not to be lightly discarded.

In the next section, therefore, we begin our construction of a mathematical model for chance. It turns out that we can make great progress by using the simple fact that our ideas about probability are closely linked to the familiar mathematical ideas of proportion and ratio.

Finally, we make the trivial point that, although the words chance, likelihood, probability, and so on mean much the same in everyday speech, we will only use one of these. What follows is thus a theory of probability.

0.3 Symmetry

Blind Fortune still bestows her gifts on such as cannot use them.

Ben Jonson

We begin with some basic ideas and notation. Many occurrences of probability appear in everyday statements such as:

The probability of red in (American) roulette is $\frac{18}{38}$.

The probability of a head when you flip a coin is 50%.

The probability of a spade on cutting a pack of cards is 25%.

Many other superficially different statements about probability can be reformulated to appear in the above format. This type of statement is in fact so frequent and fundamental that we use a standard abbreviation and notation for it. Anything of the form

the probability of A is p

will be written as:

$$\mathbf{P}(A) = p.$$

In many cases, p may represent an adjective denoting quantity, such as “low” or “high.” In the examples above, A and p were, respectively,

$$\begin{aligned} A &\equiv \text{red}, & p &= 18/38 \\ A &\equiv \text{heads}, & p &= 50\% \\ A &\equiv \text{spade}, & p &= 25\%. \end{aligned}$$

You can easily think of many similar statements. Our first urgent question is, where did those values for the probability p come from? To answer this, let us consider what happens when we pick a card at random from a conventional pack. There are 52 cards, of which 13 are spades. The implication of the words “at random” is that any card is equally likely to be selected, and the proportion of the pack comprising the spade suit is $13/52 = \frac{1}{4}$. Our intuitive feelings about symmetry suggest that the probability of picking a spade is directly proportional to this fraction, and by convention we choose the constant of proportionality to be unity. Hence,

$$\mathbf{P}(\text{spade}) = \frac{1}{4} = 25\%.$$

Exactly the same intuitive interpretation comes into play for any random procedure having this kind of symmetry.

Example: American Roulette These wheels have 38 compartments, of which 18 are red, 18 are black, and two are green (the zeros). If the wheel has been made with equal-size compartments (and no hidden magnets, or subtle asymmetries), then the ball has 18 chances to land in red out of the 38 available. This suggests

$$\mathbf{P}(\text{red}) = 18/38.$$



In the case of a fair coin, of course, there are only two equally likely chances to $\mathbf{P}(\text{Head}) = 50\%$ and $\mathbf{P}(\text{Tail}) = 50\%$. This particular case of equal probabilities has passed into the language in the expression a “fifty-fifty” chance (first used in print by P.G. Wodehouse in his novel *The Little Nugget*).

In general, this argument (or expression of our intuition) leads to the following definition of probability. Suppose that some procedure with a random outcome has n distinct possible outcomes, and suppose further that by symmetry (or by construction or supposition) these outcomes are equally likely. Then if A is any collection of r of these outcomes, we define

$$(1) \quad \mathbf{P}(A) = \frac{r}{n} = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}.$$

Note that in this case we must have

$$(2) \quad 0 \leq \mathbf{P}(A) \leq 1,$$

because $0 \leq r \leq n$. Furthermore, if A includes all n possible outcomes, then $\mathbf{P}(A) = 1$. At the other extreme, $\mathbf{P}(A) = 0$ if A contains none of the possible outcomes.

Here is another simple example.

(3) **Example: Die** With the probability of any event now defined by (1), it is elementary to find the probability of any of the events that may occur when we roll a die. The number shown may be (for example) even, odd, prime or perfect, and we denote these events by A , B , C and D respectively. Here $n = 6$, and for $A = \{2 \text{ or } 4 \text{ or } 6\}$ we have $r = 3$. The probability that it shows an even number is

$$\mathbf{P}(A) = \mathbf{P}(\{2 \text{ or } 4 \text{ or } 6\}) = \frac{3}{6} = \frac{1}{2}.$$

Likewise, and equally trivially, we find that

$$\mathbf{P}(\text{odd}) = \mathbf{P}(B) = \mathbf{P}(\{1 \text{ or } 3 \text{ or } 5\}) = \frac{1}{2}$$

$$\mathbf{P}(\text{prime}) = \mathbf{P}(C) = \mathbf{P}(\{2 \text{ or } 3 \text{ or } 5\}) = \frac{1}{2}$$

$$\mathbf{P}(\text{perfect}) = \mathbf{P}(D) = \mathbf{P}(\{6\}) = \frac{1}{6}.$$

These values of the probabilities are not inconsistent with our ideas about how the symmetries of this die should express themselves when it is rolled. ●

This idea or interpretation of probability is very appealing to our common intuition. It is first found nascent in a poem entitled “De Vetula,” which was widely distributed in manuscript form from around 1250 onward. It is, of course, extremely likely that this idea of probability had been widespread for many years before then. In succeeding years, most probability calculations during the Renaissance and the ensuing scientific revolution take this framework for granted.

However, there are several unsatisfactory features of this definition: first, there are plenty of random procedures with no discernible symmetry in the outcomes; and second, it is worrying that we do not need actually to roll a die to say that chance of a six is $\frac{1}{6}$. Surely,

actual experiments should play some part in shaping and verifying our theories about the physical world? We address this difficulty in the next section.

0.4 The Long Run

Nothing is more certain than incertainties
 Fortune is full of fresh variety
 Constant in nothing but inconstancy
Richard Barnfield

Suppose that some random procedure has several possible outcomes that are not necessarily equally likely. How can we define the probability $\mathbf{P}(A)$ of any eventuality A of interest? For example, suppose the procedure is the rolling of a die that is suspected to be weighted, or even clearly asymmetrical, in not being a perfect cube. What now is the probability of a six?

There is no symmetry in the die to help us, but we can introduce symmetry another way. Suppose you roll the die a large number n of times, and let $r(n)$ be the number of sixes shown. Then (provided the rolls were made under similar conditions) the symmetry between the rolls suggests that (at least approximately)

$$\mathbf{P}(\text{six}) = \frac{r(n)}{n} = \frac{\text{number of sixes}}{\text{number of rolls}}.$$

Furthermore, if you actually obtain an imperfect or weighted die and roll it many times, you will find that as n increases the ratio $\frac{r(n)}{n}$ always appears to be settling down around some asymptotic value. This provides further support for our taking $\frac{r(n)}{n}$ as an approximation to $\mathbf{P}(\text{six})$.

Of course, this procedure can only ever supply an approximation to the probability in question, as the ratio $r(n)/n$ changes with n . This is the sort of price that we usually pay when substituting empiricism for abstraction. There are other possible eventualities that may also confuse the issue; for example, if told that a coin, in 1 million flips, showed 500,505 heads and 499,495 tails, you would probably accept it as fair, and you would set $\mathbf{P}(\text{head}) = \frac{1}{2}$. But suppose you were further informed that all the heads formed a run preceding all the tails; would you now be quite so confident? Such a sequence might occur, but our intuition tells us that it is so unlikely as to be irrelevant to this discussion. In fact, routine gaming and other experience bears out our intuition in the long run. That is why it is our intuition; it relies not only on our own experience, but also on our gambling predecessors. You can believe in the symmetry and long-run interpretations for chances in roulette without ever having spun the wheel or wagered on it (the author has done neither).

This idea of the long run can clearly be extended to any random procedure that it is possible to repeat an arbitrary number n of times under essentially identical conditions. If A is some possible result and A occurs on $r(n)$ occasions in n such repetitions, then we say that

$$(1) \quad \mathbf{P}(A) \approx r(n)/n.$$

Thus, evaluating the ratio $r(n)/n$ offers a way of measuring or estimating the probability $\mathbf{P}(A)$ of the event A . This fact was familiar to gamblers in the Renaissance, and presumably well before then. Cardano in his *Book on Games of Chance* (written around 1520) observes that “every die, even if it is acceptable, has its favoured side.” It may be assumed that gamblers noticed that in the long run small biases in even a well-made die will be revealed in the empirical proportions of successes for the six faces.

A similar empirical observation was recorded by John Graunt in his book *Natural and Political Observations Made Upon the Bills of Mortality* (1662). He found that in a large number of births, the proportion of boys born was approximately $\frac{14}{27}$. This came as something of a surprise at the time, leading to an extensive debate. We simply interpret the observation in this statement: the probability of an as yet unborn child being male is approximately $\frac{14}{27}$. (Note that this empirical ratio varies slightly from place to place and time to time, but it always exceeds $\frac{1}{2}$ in the long run.)

Once again, we stress the most important point that $0 \leq r(n)/n \leq 1$, because $0 \leq r(n) \leq n$. It follows that the expression (1) always supplies a probability in $[0, 1]$. Furthermore, if A is impossible (and hence never occurs), $r(n)/n = 0$. Conversely, if A is certain (and hence always occurs), $r(n)/n = 1$.

0.5 Pay-Offs

Probability is expectation, founded upon partial knowledge.

George Boole

In reading the previous two sections, the alert reader will have already made the mental reservation that many random procedures are neither symmetrical nor repeatable. Classic examples include horse races, football matches, and elections. Nevertheless, bookmakers and gamblers seem to have no qualms about quoting betting odds, which are essentially linked to probabilities. (We explore this connection in more depth in Sections 0.7 and 1.6.) How is it possible to define probabilities in these contexts?

One possible approach is based on our idea of a “fair value” of a bet, which in turn is linked to the concept of mathematical averages in many cases. An illustrative example of great antiquity, and very familiar to many probabilists, is provided by the following problem.

A parsimonious innkeeper empties the last three glasses of beer (worth 13 each) from one barrel and the last two glasses (worth 8 each) from another, and mixes them in a jug holding five glasses. What is the “fair” price for a glass from the jug? The innkeeper calculates the value of the beer in the jug as $3 \times 13 + 2 \times 8 = 55$, and divides by 5 to obtain his “fair” price of 11 each, although discriminating customers may well have another view about that.

A similar idea extends to random situations. For example, suppose the benevolent but eccentric uncle of Jack and Tom flips a fair coin; if it shows heads then Jack gets \$1, if it shows tails then Tom gets \$1. If this is worth \$ p to Jack, it is worth the same to Tom. Because the uncle certainly parts with \$1, we have $2p = 1$, which is to say that a “fair” price for Jack to sell his share of the procedure *before* the uncle flips the coin is \$ μ , where

$$\mu = \frac{1}{2}.$$

More generally, if the prize at stake is $\$d$, then a fair price for either of their expected gifts is $\$d/2$. More generally still, by a similar argument, if you are to receive $\$d$ with probability p , then a fair price for you to sell this uncertain reward *in advance* of the experiment is $\$dp$.

But this argument has a converse. Suppose you are to receive $\$1$ if some event A occurs and nothing otherwise. Further suppose that you and Tom agree that he will give you $\$p$ before the experiment occurs, and that he will get your reward whatever it may be. Then, in effect, you and Tom have estimated and agreed that (at least approximately)

$$p = \mathbf{P}(A).$$

To see this, note that if you believed $\mathbf{P}(A)$ were larger than p you would hold out for a higher price, and Tom would not pay as much as p if he believed $\mathbf{P}(A)$ were less than p .

Thus, probabilities can be defined, at least implicitly, whenever people can agree on a fair price. Note that $0 \leq p \leq 1$ in every case. Finally, we observe that the idea of fair price can be extended and turns out to be of great importance in later work. For example, suppose you roll a fair five-sided die; if faces 1, 2, or 3 turn up you win $\$13$, if faces 4 or 5 turn up you win $\$8$. What is a fair price for one roll? Essentially the same arithmetic and ideas as we gave above for the parsimonious innkeeper reveals the fair price (or value) of one roll to be $\$11$ because

$$13 \times \frac{3}{5} + 8 \times \frac{2}{5} = 11.$$

We will meet this concept again under the name “expectation.”

0.6 Introspection

Probability is a feeling of the mind.

Augustus de Morgan

In the previous three sections, we defined probabilities, at least approximately, in situations with symmetry, or when we could repeat a random procedure under essentially the same conditions, or when there was a plausible agreed “fair price” for a resulting prize. However, there are many asymmetric and non-repeatable random events, and for many of these we would feel distinctly unsatisfied, or even unhappy, with the idea that their probabilities should be determined by gamblers opening a book on the question. For example, what is the probability that some accused person is guilty of the charge? Or the probability of life on another planet? Or the probability of you catching a cold this week? Or the probability that Shakespeare wrote some given sonnet of doubtful provenance? Or the probability that a picture called “Sunflowers” is by van Gogh? Or the probability that Riemann’s hypothesis is true? (It asserts that all the nontrivial zeros of a certain function have real part $+\frac{1}{2}$.) Or the probability that π^e is irrational? In this last question, would the knowledge that e^π is irrational affect the probability in your judgment?

In these questions, our earlier methods seem more or less unsatisfactory, and, indeed, in a court of law you are forbidden to use any such ideas in deciding the probability of guilt of an accused. One is led to the concept of probability as a “degree of belief.”

If we assign a probability to any of these eventualities, then the result must of necessity be a personal or subjective assessment. Your figure need not be the same as mine or anyone else's, and any probability so obtained is called subjective or personal.

In fact, there is a strong (although perhaps minority) body of opinion that maintains that all probabilities are subjective (cf the remark of A. de Morgan at the head of this section). They argue that appeals to symmetry, the long run, or fair value, merely add a spurious objectivity to what is essentially intuition, based on personal experience, logical argument, and experiments (where these are relevant). The examples above are then simply rather trivial special cases of this general definition.

According to this approach, a probability is a measure of your "degree of belief" in the guilt of the accused, the truth of some assertion, or that a die will show a six. In the classical case based on dice, etc., this belief rests on symmetry, as it does in the "long-run" relative-frequency interpretation. The "fair price" approach uses the beliefs of all those concerned in fixing such a price. Furthermore, the approach via degrees of belief allows (at least in principle) the possibility that such probabilities could be determined by strictly logical statements relating what we know for sure to the uncertain eventuality in question. This would be a kind of inductive probability logic, an idea that was first suggested by Leibniz and later taken up by Boole. During the past century, there have been numerous clever and intriguing books about various approaches to establishing such an axiomatic framework.

This argument is clearly seductive, and for all the above reasons it is tempting to regard all types of probability as a "feeling of the mind," or as a "degree of belief." However, there are several drawbacks. First, in practice, it does not offer a wholly convincing and universally accepted way of defining or measuring probability, except in the cases discussed above. Thus, the alleged generality is a little artificial because different minds feel differently.

Second, setting this more general idea in a formal framework requires a great deal more effort and notation, which is undesirable for a first approach. Finally, it is in any case necessary for practical purposes to arrange things so that the rules are the same as those obtained from the simpler arguments that we have already outlined.

For these reasons, we do not further pursue the dream of a universal interpretation of probability; instead, we simply note this remark of William Feller:

All definitions of probability fall short of the actual practice.

0.7 FAQs

Neither physicists nor philosophers can give any convincing account of what "physical reality" is.

G.H. Hardy

"What is the meaning of it, Watson?" said Holmes, solemnly, as he laid down the paper. . . . "It must tend to some end, or else our universe is ruled by chance, which is unthinkable."

A. Conan Doyle, The Adventure of the Cardboard Box

In the preceding sections, we agree to develop a mathematical theory of probability and discuss some interpretations of probability. These supply definitions of probability in

some simple cases and, in any case, we interpret probability as an extension of the idea of proportion.

The net result of this preliminary reconnaissance is our convention (or rule) that probability is a number between zero and unity inclusive, where impossibility corresponds to 0, and certainty corresponds to 1.

In Chapter 1, we probe more deeply into our ideas about probability, to discover more subtle and important rules that describe its behaviour, and generate our mathematical model. However, there are a few very common doubts and queries that it is convenient to dispose of here.

- (1) The first, and perhaps in one way, key question is to ask if anything is really random? This is called the question of determinism, and its importance was realised early on. One point of view is neatly expressed by Laplace in these two extracts from his monumental work on probability:

The path of a molecule is fixed as certainly as the orbits of the planets: the only difference between the two is due to our ignorance. Probability relates partly to this ignorance and partly to our knowledge. . . . Thus, given a sufficiently great intelligence that could encompass all the forces of nature and the details of every part of the universe, . . . nothing would be uncertain and the future (as well as the past) would be present before its eyes.

This is called “determinism” because it asserts that the entire future is determined by a complete knowledge of the past and present. There are many problems with this, not the least of them being the unattractive corollary that people do not have free will. Your every action is inevitable and unavoidable. However, no one actually believes this, unless wasting time on fruitless cerebration is also unavoidable, and so on. The difficulties are clear; fortunately, on one interpretation of quantum theory, certain activities of elementary particles are genuinely unpredictable, an idea expressed by Heisenberg’s uncertainty principle. But this view also leads to paradoxes and contradictions, not the least of which is the fact that some events must perforce occur with no cause. However, causeless events are about as unattractive to the human mind as the nonexistence of free will, as expressed by Einstein’s remark, “In any event, I am sure that God does not dice.” At this point we abandon this discussion, leaving readers to ponder it for as long as they desire. Note, first, that however long you ponder it, you will not produce a resolution of the problems and, second, that none of this matters to our theory of probability.

- (2) If you accept that things can be random, the second natural question is to ask: What is probability really? The answer to this is that our intuition of probability begins as a large portmanteau of empirical observations that the universe is an uncertain place, with many areas of doubt and unpredictability. In response to this, we form a theory of probability, which is a description of what we observe in terms of (mathematical) rules. In the sense in which the question is usually posed, we here neither know nor care what probability “really” is. (If we were engaged in a certain type of philosophical investigation, this question might concern us, but we are not.) It is the same in all sciences; we label certain concepts such as “mass,” “light,” “particle,” and describe their properties by mathematical rules. We may never know what light or mass “really” are, nor does it matter. Furthermore, for example, even if it turned out that mass is

supplied by the Higgs boson, the next question would be to ask what is a boson “really”? You might just as well ask what are numbers really?

- (3) When it is appreciated that probability is our constructed model, it is next natural to ask is it the only one or are there others?

There are indeed many other theories devised to explain uncertain phenomena. For example, the quantum theory designed to explain the odd behaviour of elementary particles uses a probabilistic structure quite different from that seen here. Another area in which a different theoretical structure would be necessary is in applying probability to questions of legal proof “beyond reasonable doubt” or “on the balance of probabilities.” The answer in this case is still a matter for debate.

At the other extreme, attempts have been made to construct more general theories with a smaller set of rules. So far, none of these is in general use.

- (4) Accepting that we need a theory, another natural question is; Does it have to be so abstract and complicated? The answer here is (of course) yes, but the reason is not just the perversity of professors. The first point is that probability is not directly tangible in the same way as the raw material of many branches of physics. Electricity, heat, magnetism, and so on, will all register on the meters of appropriate instruments. There is no meter to record the presence of probability, except us. But we are imperfect instruments, and not infrequently rather confused. It is the case that most people’s intuition about problems in chance will often lead them grossly astray, even with very simple concepts. Although many examples appear later, we mention a few here:

- (a) *The base rate fallacy.* This appears in many contexts, but it is convenient to display it in the framework of a medical test for a disease that affects one person in 100,000.

You have a test for the disease that is 99% accurate. (That is to say, when applied to a sufferer, it shows positive with probability 99%; when applied to a nonsufferer, it shows negative with probability 99%.) What is the probability that you have the disease if your test shows a positive result? Most people’s untutored intuition would lead them to think the chance is high, or at least not small.

In fact, the chance of having the disease, given the positive result, is less than one in a 1,000; indeed, it is more likely that the test was wrong.

- (b) *The Monty Hall problem.* This is now so well-known as to hardly need stating, but for all our extraterrestrial readers here it is: You are a contestant in a game show. A nice car and two feral goats are randomly disposed behind three doors, one to each door. You choose a door to obtain the object it conceals. The presenter does not open your chosen door, but opens another door that turns out to reveal a goat. Then the presenter offers you the chance to switch your choice to the final door. Do you gain by so doing? That is to say, what is the probability that the final door conceals the car?

Many people’s intuition tells them that, given the open door, the car is equally likely to be behind the remaining two; so there is nothing to be gained by switching. In fact, this is almost always wrong; you should switch. However, even this simple problem raises issues of quite surprising complexity, which require our sophisticated theory of probability for their resolution.

- (c) *Coincidences.* Twenty-three randomly selected people are listening to a lecture on chance. What is the probability that at least two of them were born on the same

day of the year? Again, untutored intuition leads most people to guess that the chances of this are rather small. In fact, in a random group of 23 people, it is more likely than not that at least two of them were born on the same day of the year.

This list of counterintuitive results could be extended indefinitely, but this should at least be enough to demonstrate that only mathematics can save people from their flawed intuition with regard to chance events.

- (5) A very natural FAQ that requires an answer is: What have “odds” got to do with probability? This is an important question, because there is a longstanding link between chance and betting, and bookmakers quote their payouts in terms of these “odds.” The first key point is that there are two kinds of odds: fair odds and pay-off odds.
- (a) *Fair odds*: If the occurrence of some event is denoted by A , then A^c denotes the nonoccurrence of A , (\equiv the event that A does not occur). If the probability of A is p and the probability of A^c is q , then the odds against A are $q : p$; (pronounced q to p). The odds on A are $p : q$. In the case of n equally likely outcomes of which r yield A and $n - r$ yield A^c , the odds on A are $r : n - r$. Thus, the odds against a six when rolling a fair die are 5 to 1, and so on.
- (b) *Pay-off odds*. These are the odds that are actually offered by bookmakers and casinos. In a sense, they could be called Unfair Odds because they are fixed to ensure that the advantage lies with the casino. For example, suppose you bet on an even number arising in an American roulette wheel. We have noted above that the probability $\mathbf{P}(E)$ of an even number is $\frac{18}{38} = \frac{9}{19}$, so the fair odds on an even number are 9 : 10; the fair odds against are 10 : 9. However, the pay-off odds in the casino are 1 : 1, which is to say that you get your stake and the same again if you win. The fair value of this \$1 bet (as discussed in section 0.5) is therefore

$$\frac{9}{19} \times 2 = \frac{18}{19} = 1 - \frac{1}{19}.$$

The fact that this is less than your stake is an indication that these odds are stacked against you.

Of course, if the casino paid out at the fair odds of 10 : 9, the value of a bet of \$1 is

$$\frac{9}{19} \left(\frac{10}{9} + 1 \right) = 1,$$

which is equal to your stake, and this is to be regarded as fair.

The difference between the fair odds (which reflect the actual probabilities) and the pay-off odds is what guarantees the casino’s profit in the long run.

We conclude with a couple of less problematical questions that arise in more everyday circumstances.

- (6) “The weather forecast gave only a 5% chance of rain today, but then it did rain. What is the use of that?”

One sympathizes with this question, but of course the whole point of probability is to discuss uncertain eventualities in advance of their resolution. The theory cannot hope to turn uncertainty into certainty; instead, it offers the prospect of discussing

these chances rationally, with the intention of making better-informed decisions as a result. It is always better to play with the odds than against them.

- (7) Will a knowledge of probability enable me to design a betting system to win at roulette? Absolutely not! [unless, of course, your betting system comprises the one simple rule: “Buy a casino.”]

0.8 History

Errors using inadequate data are much less than those using no data at all.

Charles Babbage

The excitement that a gambler feels when making a bet is equal to the amount he might win multiplied by the chance of winning it.

Blaise Pascal

One of the most interesting features of the history of probability is how remarkably short it is. The Ancient Greeks and others had been prompted (by everyday problems and their curiosity) to develop an extensive knowledge of geometry, astronomy, and numbers. Other branches of mathematics made great strides during the Italian Renaissance. Nevertheless, it was left to Pascal, Fermat, and Huygens to inaugurate mathematical probability in 1654 to 1656, with their work on the Problem of the Points (which concerned the fair division of the prize or stakes in unfinished games of chance), and the Gamblers Ruin Problem (see Examples 2.11 and 4.12).

This tardy start is all the more surprising when one notes that games of chance, using dice and other familiar randomisers, were certainly widespread throughout the classical era, and were almost certainly common before that period.

For example, the game known as *Alea* was widespread throughout the Roman Empire at all levels of society. At various times, laws proscribed or restricted it; Marcus Aurelius was so addicted to the game that he had a personal croupier to roll his dice. The 6×6 squared boards, often inscribed appropriately, are found across the imperial domain. Various numbers of dice were used, usually three; the most desirable outcome was called the Venus throw, the least desirable was the Dogs. Note that the Latin for die is *alea* (from which we get aleatory meaning random), and the game used counters, from which the word *calculus* has also entered our language. The game was also played by the Greeks, with the legend that it was invented by Palamedes to occupy the tedious years of the Trojan wars, particularly the celebrated siege. (Another legend attributes it to the Lydians, who used it to distract their minds from hunger during famine.) The Greeks called the game *pessoi*, after the counters, a die was *tessera*. It seems safe to say that the origins of the game are lost in early antiquity.

Despite these ancient origins, there is no evidence that anyone attempted to calculate chances, and play accordingly, until the second millennium. A poem entitled “*De Vetula*,” dated to around 1250 AD, includes elementary calculations of chances involving dice. Similar ideas and calculations can be found sporadically throughout Europe over the next four centuries, but the spirit of the times and difficulties of communication hampered any

serious development. Nevertheless, we find Cardano writing a book around 1520, entitled *On Games of Chance*, in which the ideas of symmetry and long-run frequency are nearly clear. Galileo was certainly taking such notions for granted in the early 17th century when he calculated the odds on rolling 9 or 10, respectively, with three dice.

Following the ground-breaking correspondence of Pascal and Fermat on the problem of the points (which had been known but unsolved for two centuries previously), there was a comparatively swift and substantial development. Christiaan Huygens (1657) wrote a book on numerous problems in probability, followed by the books of James Bernoulli (1713), Pierre de Montmort (1708), and Abraham de Moivre (1718, 1738, 1756).

The final edition of de Moivre's book includes a law of large numbers, a central limit theorem (see Sections 7.5.11 and 8.9), and quite sophisticated solutions of problems using generating functions (see Section 3.6 and Chapter 6). The pace of development accelerated during the 19th century, until the structure that is most in use today was rigorously codified by A. Kolmogorov in his book of 1933, *Grundbegriffe der Wahrscheinlichkeitsrechnung*. (We note, however, that other systems have been, are being, and will be used to model probability.)

Review In this chapter, we discuss our intuitive ideas about chance and suggest how they can help us to construct a more formal theory of probability. In particular, we exploit the interpretations of chances as a simple proportion in situations with symmetry and as relative frequency in the long run. These suggest that:

- Any probability should be a number between zero and one, inclusive.
- Things that are impossible should have zero probability.
- Things that are certain should have probability one.

Thus, we can picture probabilities as lying on a scale between zero and one, where the more unlikely eventualities have their probabilities nearer to zero and the more likely eventualities have their probabilities nearer to one. The following chapters use similar arguments to develop more complicated rules and properties of probability.

Appendix: Review of Elementary Mathematical Prerequisites

It is difficult to make progress in any branch of mathematics without using the ideas and notation of sets and functions. Indeed, it would be perverse to try to do so because these ideas and notation are helpful in guiding our intuition and solving problems. (Conversely, almost the whole of mathematics can be constructed from these few simple concepts.) We therefore give a brief synopsis of what we need here for completeness, although it is likely that the reader will already be familiar with this.

Notation

We use a good deal of familiar standard mathematical notation in this book. The basic notation for sets and functions is set out below. More specialized notation for probability theory is introduced as required throughout the book, and recorded in Chapter Reviews and the Index of Notation.

We take this opportunity to list some fundamental notation that you are likely to see soon:

e	the base of natural logarithms; Euler's number
$\log x$	the logarithm of x to base e , unless otherwise stated
$\log_2 x$	the logarithm of x to base 2
π	the ratio of circumference to diameter for a circle
Ω	the universal set
$n!$	$n(n-1)\dots 3.2.1$; factorial n
$ x $	modulus or absolute value of x
$[x]$	the integer part of x
\mathbb{R}	the real line
\mathbb{Z}	the integers
$x \wedge y = \min\{x, y\}$	the smaller of x and y
$x \vee y = \max\{x, y\}$	the larger of x and y
$x^+ = x \vee 0$	
$\sum_{r=1}^n a_r = a_1 + \dots + a_n$	summation symbol
$\prod_{r=1}^n a_r = a_1 a_2 \dots a_n$	product symbol

Sets

A *set* is a collection of things that are called the elements of the set. The elements can be any kind of entity: numbers, people, poems, blueberries, points, lines, and so on, endlessly.

For clarity, upper case letters are always used to denote sets. If the set S includes some element denoted by x , then we say x belongs to S and write $x \in S$. If x does not belong to S , then we write $x \notin S$.

There are essentially two ways of defining a set, either by a *list* or by a *rule*.

Example If S is the set of numbers shown by a conventional die, then the *rule* is that S comprises the integers lying between 1 and 6 inclusive. This may be written formally as follows:

$$S = \{x : 1 \leq x \leq 6 \text{ and } x \text{ is an integer}\}.$$

Alternatively, S may be given as a *list*:

$$S = \{1, 2, 3, 4, 5, 6\}. \quad \bullet$$

One important special case arises when the rule is impossible; for example, consider the set of elephants playing football on Mars. This is impossible (there is no pitch on Mars) and the set therefore is empty; we denote the empty set by ϕ . We may write ϕ as $\{\}$.

If S and T are two sets such that every element of S is also an element of T , then we say that T includes S and write either $S \subseteq T$ or $S \subset T$. If $S \subset T$ and $T \subset S$, then S and T are said to be equal and we write $S = T$.

Note that $\phi \subset S$ for every S . Note also that some books use the symbol “ \subseteq ” to denote inclusion and reserve “ \subset ” to denote strict inclusion, that is to say, $S \subset T$ if every element of S is in T and some element of T is not in S . We do not make this distinction.

Combining Sets

Given any nonempty set, we can divide it up, and given any two sets, we can join them together. These simple observations are important enough to warrant definitions and notation.

Definition *Let A and B be sets. Their union, denoted by $A \cup B$, is the set of elements that are in A or B , or in both. Their intersection, denoted by $A \cap B$, is the set of elements in both A and B .* ▲

Note that in other books the union may be referred to as the *join* or *sum*; the intersection may be referred to as the *meet* or *product*. We do not use these terms. Note the following.

Definition *If $A \cap B = \phi$, then A and B are said to be disjoint.* ▲

We can also remove bits of sets, giving rise to set differences, as follows.

Definition *Let A and B be sets. That part of A that is not also in B is denoted by $A \setminus B$, called the difference of A from B . Elements that are in A or B but not both, comprise the symmetric difference, denoted by $A \Delta B$.* ▲

Finally, we can combine sets in a more complicated way by taking elements in pairs, one from each set.

Definition *Let A and B be sets, and let*

$$C = \{(a, b) : a \in A, b \in B\}$$

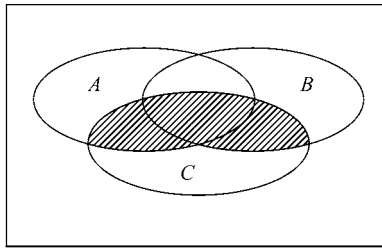
be the set of ordered pairs of elements from A and B . Then C is called the product of A and B and denoted by $A \times B$. ▲

Example *Let A be the interval $[0, a]$ of the x -axis, and B the interval $[0, b]$ of the y -axis. Then $C = A \times B$ is the rectangle of base a and height b with its lower left vertex at the origin, when $a, b > 0$.* ●

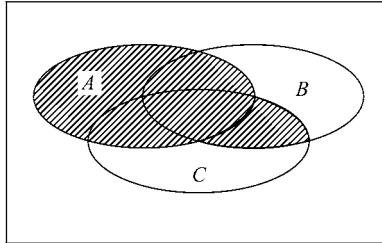
Venn Diagrams

The above ideas are attractively and simply expressed in terms of *Venn diagrams*. These provide very expressive pictures, which are often so clear that they make algebra redundant (see Figure 0.1).

In probability problems, all sets of interest A lie in a universal set Ω , so that $A \subset \Omega$ for all A . That part of Ω that is not in A is called the *complement* of A , denoted by A^c .



The shaded area is $(A \cup B) \cap C$ and $(A \cap C) \cup (B \cap C)$.



The shaded area is $A \cup (B \cap C)$ and $(A \cup B) \cap (A \cup C)$.

Figure 0.1 Venn diagrams.

Formally,

$$A^c = \Omega \setminus A = \{x : x \in \Omega, x \notin A\}.$$

Obviously, from the diagram or by consideration of the elements

$$A \cup A^c = \Omega, \quad A \cap A^c = \phi, \quad (A^c)^c = A.$$

Clearly, $A \cap B = B \cap A$ and $A \cup B = B \cup A$, but we must be careful when making more intricate combinations of larger numbers of sets. For example, we cannot write down simply $A \cup B \cap C$; this is not well defined because it is not always true that

$$(A \cup B) \cap C = A \cup (B \cap C).$$

We use the obvious notation

$$\bigcup_{r=1}^n A_r = A_1 \cup A_2 \cup \cdots \cup A_n,$$

$$\bigcap_{r=1}^n A_r = A_1 \cap A_2 \cap \cdots \cap A_n.$$

Definition If $A_j \cap A_k = \phi$ for $j \neq k$ and

$$\bigcup_{r=1}^n A_r = \Omega,$$

then the collection $(A_r; 1 \leq r \leq n)$ is said to form a partition of Ω . ▲

Size

When sets are countable, it is often useful to consider the number of elements they contain; this is called their *size* or *cardinality*. For any set A , we denote its size by $|A|$; when sets have a finite number of elements, it is easy to see that size has the following properties.

If sets A and B are disjoint, then

$$|A \cup B| = |A| + |B|,$$

and more generally, when A and B are not necessarily disjoint,

$$|A \cup B| + |A \cap B| = |A| + |B|.$$

Naturally, $|\phi| = 0$, and if $A \subseteq B$, then

$$|A| \leq |B|.$$

Finally, for the product of two such finite sets $A \times B$, we have

$$|A \times B| = |A| \times |B|.$$

When sets are infinite or uncountable, a great deal more care and subtlety is required in dealing with the idea of size. However, we intuitively see that we can consider the length of subsets of a line, or areas of sets in a plane, or volumes in space, and so on. It is easy to see that if A and B are two subsets of a line, with lengths $|A|$ and $|B|$, respectively, then in general

$$|A \cup B| + |A \cap B| = |A| + |B|.$$

Therefore $|A \cup B| = |A| + |B|$ when $A \cap B = \phi$.

We can define the product of two such sets as a set in the plane with area $|A \times B|$, which satisfies the well-known elementary rule for areas and lengths

$$|A \times B| = |A| \times |B|$$

and is thus consistent with the finite case above. Volumes and sets in higher dimensions satisfy similar rules.

Functions

Suppose we have sets A and B , and a rule that assigns to each element a in A a unique element b in B . Then this rule is said to define a *function* from A to B ; for the corresponding elements, we write

$$b = f(a).$$

Here the symbol $f(\cdot)$ denotes the rule or function; often we just call it f . The set A is called the domain of f , and the set of elements in B that can be written as $f(a)$ for some a is called the range of f ; we may denote the range by R .

Anyone who has a calculator is familiar with the idea of a function. For any function key, the calculator will supply $f(x)$ if x is in the domain of the function; otherwise, it says "error."

Inverse Function

If f is a function from A to B , we can look at any b in the range R of f and see how it arose from A . This defines a rule assigning elements of A to each element of R , so if the rule assigns a unique element a to each b this defines a function from R to A . It is called the *inverse function* and is denoted by $f^{-1}(\cdot)$:

$$a = f^{-1}(b).$$

Example: Indicator Function Let $A \subset \Omega$ and define the following function $I(\cdot)$ on Ω :

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

Then I is a function from Ω to $\{0, 1\}$; it is called the *indicator* of A because by taking the value 1 it indicates that $\omega \in A$. Otherwise, it is zero. ●

This is about as simple a function as you can imagine, but it is surprisingly useful. For example, note that if A is finite you can find its size by summing $I(\omega)$ over all ω :

$$|A| = \sum_{\omega \in \Omega} I(\omega).$$

Series and Sums

Another method I have made use of, is that of Infinite Series, which in many cases will solve the Problems of Chance more naturally than Combinations.

A. de Moivre, Doctrine of Chances, 1717

What was true for de Moivre is equally true today, and this is therefore a convenient moment to remind the reader of some general and particular properties of series. Note that $n! = n(n-1) \dots 3 \cdot 2 \cdot 1$, and that $0! = 1$, by convention. Also,

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

and

$$M(a, b, c) = \frac{(a+b+c)!}{a!b!c!}.$$

Finite Series

Consider the series

$$s_n = \sum_{r=1}^n a_r = a_1 + a_2 + \dots + a_n.$$

The variable r is a *dummy variable* or *index of summation*, so any symbol will suffice:

$$\sum_{r=1}^n a_r \equiv \sum_{i=1}^n a_i.$$

In general,

$$\sum_{r=1}^n (ax_r + by_r) = a \sum_{r=1}^n x_r + b \sum_{r=1}^n y_r.$$

In particular,

$$\sum_{r=1}^n 1 = n;$$

$$\sum_{r=1}^n r = \frac{1}{2}n(n+1), \quad \text{the arithmetic sum;}$$

$$\sum_{r=1}^n r^2 = \frac{1}{6}n(n+1)(2n+1) = 2\binom{n+1}{3} + \binom{n+1}{2};$$

$$\sum_{r=1}^n r^3 = \left(\sum_{r=1}^n r\right)^2 = \frac{1}{4}n^2(n+1)^2;$$

$$\sum_{r=0}^n \binom{n}{r} x^r y^{n-r} = (x+y)^n, \quad \text{the binomial theorem;}$$

$$\sum_{\substack{a+b+c=n \\ a,b,c \geq 0}} M(a,b,c)x^a y^b z^c = \sum_{\substack{a+b+c=n \\ a,b,c \geq 0}} \binom{a+b+c}{a+b} \binom{a+b}{a} x^a y^b z^c$$

$$= (x+y+z)^n, \quad \text{the multinomial theorem;}$$

$$\sum_{r=0}^n x^r = \frac{1-x^{n+1}}{1-x}, \quad \text{the geometric sum.}$$

Limits

Often, we have to deal with infinite series. A fundamental and extremely useful concept in this context is that of the limit of a sequence.

Definition Let $(s_n; n \geq 1)$ be a sequence of real numbers. If there is a number s such that $|s_n - s|$ may ultimately always be as small as we please, then s is said to be the limit of the sequence s_n . Formally, we write

$$\lim_{n \rightarrow \infty} s_n = s$$

if and only if for any $\varepsilon > 0$, there is a finite n_0 such that $|s_n - s| < \varepsilon$ for all $n > n_0$. ▲

Notice that s_n need never actually take the value s , it must just get closer to it in the long run (e.g., let $x_n = n^{-1}$).

Infinite Series

Let $(a_r; r \geq 1)$ be a sequence of terms, with partial sums

$$s_n = \sum_{r=1}^n a_r, \quad n \geq 1.$$

If s_n has a finite limit s as $n \rightarrow \infty$, then the sum $\sum_{r=1}^{\infty} a_r$ is said to *converge* with sum s . Otherwise, it *diverges*. If $\sum_{r=1}^{\infty} |a_r|$ converges, then $\sum_{r=1}^{\infty} a_r$ is said to be *absolutely convergent*.

For example, in the geometric sum in I above, if $|x| < 1$, then $|x|^n \rightarrow 0$ as $n \rightarrow \infty$. Hence,

$$\sum_{r=0}^{\infty} x^r = \frac{1}{1-x}, \quad |x| < 1,$$

and the series is absolutely convergent for $|x| < 1$. In particular, we have the *negative binomial theorem*:

$$\sum_{r=0}^{\infty} \binom{n+r-1}{r} x^r = (1-x)^{-n}.$$

This is true even when n is not an integer. For example,

$$\begin{aligned} (1-x)^{-1/2} &= \sum_{r=0}^{\infty} \binom{r-\frac{1}{2}}{r} x^r = \sum_{r=0}^{\infty} \left(r - \frac{1}{2}\right) \\ &\quad \times \left(r - \frac{1}{2} - 1\right) \dots \left(r - \frac{1}{2} - (r-1)\right) x^r / r! \\ &= 1 + \frac{1}{2}x + \frac{3}{2} \times \frac{1}{2} \times \frac{x^2}{2!} + \frac{5}{2} \times \frac{3}{2} \times \frac{1}{2} \times \frac{x^3}{3!} + \dots \\ &= \sum_{r=0}^{\infty} \binom{2r}{r} \left(\frac{x}{4}\right)^r. \end{aligned}$$

In particular, we often use the case $n = 2$:

$$\sum_{r=0}^{\infty} (r+1)x^r = (1-x)^{-2}.$$

Also, by definition, for all x , where e is the base of natural logarithms,

$$\exp x = e^x = \sum_{r=0}^{\infty} \frac{x^r}{r!}$$

and, for $|x| < 1$,

$$-\log(1-x) = \sum_{r=1}^{\infty} \frac{x^r}{r}.$$

An important property of e^x is the *exponential limit theorem*:

$$\text{as } n \rightarrow \infty, \left(1 + \frac{x}{n}\right)^n \rightarrow e^x.$$

This has a useful generalization: let $r(n, x)$ be any function such that $nr(n, x) \rightarrow 0$ as $n \rightarrow \infty$, then

$$\left\{1 + \frac{x}{n} + r(n, x)\right\}^n \rightarrow e^x, \text{ as } n \rightarrow \infty.$$

Finally, note that we occasionally use special identities such as

$$\sum_{r=1}^{\infty} \frac{1}{r^2} = \frac{\pi^2}{6} \quad \text{and} \quad \sum_{r=1}^{\infty} \frac{1}{r^4} = \frac{\pi^4}{90}.$$

1

Probability

And of all axioms this shall win the prize.
'Tis better to be fortunate than wise.

John Webster

Men's judgements are a parcel of their fortunes.
W. Shakespeare, Antony and Cleopatra

1.1 Notation and Experiments

In the course of everyday life, we become familiar with chance and probability in various contexts. We express our ideas and assessments in many ways, such as:

It will very likely rain.

It is almost impossible to hole this putt.

That battery may work for a few more hours.

Someone will win the lottery, but it is most unlikely to be one of us.

It is about a 50-50 chance whether share prices will rise or fall today.

You may care to amuse yourself by noting more such judgments of uncertainty in what you say and in the press.

This large range of synonyms, similes, and modes of expression may be aesthetically pleasing in speech and literature, but we need to become much more precise in our thoughts and terms. To aid clarity, we make the following.

- (1) **Definition** *Any well-defined procedure or chain of circumstances is called an experiment. The end results or occurrences are called the outcomes of the experiment. The set of possible outcomes is called the sample space (or space of outcomes) and is denoted by the Greek letter Ω .* ▲

In cases of interest, we cannot predict with certainty how the experiment will turn out, rather we can only list the collection of possible outcomes. Thus, for example:

	Experiment	Possible outcomes
(a)	Roll a die	One of the faces
(b)	Flip a coin	Head or tail

- | | |
|--------------------------|----------------------------------------------------------------|
| (c) Buy a lottery ticket | Win a prize, or not |
| (d) Deal a bridge hand | All possible arrangements of 52 cards into
four equal parts |
| (e) Run a horse race | Any ordering of the runners |

Typically, probability statements do not refer to individual outcomes in the sample space; instead, they tend to embrace collections of outcomes or subsets of Ω . Here are some examples:

Experiment	Set of outcomes of interest
(a) Deal a poker hand	Have at least a pair
(b) Buy a share option	Be in the money at the exercise date
(c) Telephone a call centre	Get through to a human in less than 1 hour
(d) Buy a car	It runs without major defects for a whole year
(e) Get married	Stay married

Clearly this is another list that you could extend without bound. The point is that in typical probability statements of the form

the probability of A is p ,

which we also write as

$$\mathbf{P}(A) = p,$$

the symbol A represents groups of outcomes of the kind exemplified above. Furthermore, we concluded in Chapter 0 that the probability p should be a number lying between 0 and 1 inclusive. A glance at the Appendix to that chapter makes it clear that $\mathbf{P}(\cdot)$ is in fact simply a function on these subsets of Ω , which takes values in $[0, 1]$. We make all this formal; thus:

(2) **Definition** *An event A is a subset of the sample space Ω .* ▲

(3) **Definition** *Probability is a function, defined on events in Ω , that takes values in $[0, 1]$. The probability of the event A is denoted by $\mathbf{P}(A)$.* ▲

(4) **Example: Two Dice** Suppose the experiment in question is rolling two dice. (Note that in this book a “die” is a cube, conventionally numbered from 1 to 6, unless otherwise stated.) Here are some events:

- (a) $A =$ Their sum is 7.
- (b) $B =$ The first die shows a larger number than the second.
- (c) $C =$ They show the same.

We may alternatively display these events as a list of their component outcomes, so

$$A = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\},$$

and so on, but this is often very tedious.

Of course this experiment has the type of symmetry we discussed in Section 0.3, so we can also compute

$$\mathbf{P}(A) = \mathbf{P}(C) = \frac{1}{6}, \quad \text{and} \quad \mathbf{P}(B) = \frac{15}{36} = \frac{5}{12}. \quad \bullet$$

(5) **Example: Darts** A dart is thrown to hit a chess board at random. Here, “at random” clearly means that it is equally likely to hit any point of the board. Possible events are:

- (a) A = It hits a white square.
- (b) B = It lands within one knight’s move of a corner square.

This problem is also symmetrical in the sense of Section 0.3, and so we calculate $\mathbf{P}(A) = \frac{1}{2}$ and $\mathbf{P}(B) = \frac{1}{8}$. \bullet

Of course, the brief definitions above raise more questions than they answer. How should the probability function behave when dealing with two or more events? How can it be extended to cope with changes in the conditions of the experiment? Many more such questions could be posed. It is clear that the brief summary above calls for much explanation and elaboration. In the next few sections, we provide a few simple rules (or axioms) that define the properties of events and their probabilities. This choice of rules is guided by our experience of real events and their likelihoods, but our experience and intuition cannot *prove* that these rules are true or say what probability “really” is. What we can say is that, starting with these rules, we can derive a theory that provides an elegant and accurate description of many random phenomena, ranging from the behaviour of queues in supermarkets to the behaviour of nuclear reactors.

1.2 Events

Let us summarise our progress to date. Suppose we are considering some experiment such as tossing a coin. To say that the experiment is well defined means that we can list all the possible outcomes. In the case of a tossed coin, the list reads: (head, tail). For a general (unspecified) experiment, any particular outcome is denoted by ω ; the collection of all outcomes is called the *sample space* and is denoted by Ω .

Any specified collection of outcomes in Ω is called an *event*. Upper case letters such as A , B , and C are used to denote events; these may have suffices or other adornments such as A_i , \bar{B} , C^* , and so on. If the outcome of the experiment is ω and $\omega \in A$, then A is said to *occur*. The set of outcomes not in A is called the complement of A and is denoted by A^c .

In particular, the event that contains all possible outcomes is the *certain event* and is denoted by Ω . Also, the event containing no outcomes is the *impossible event* and is denoted by ϕ . Obviously, $\phi = \Omega^c$. What we said in Chapter 0 makes it natural to insist that $\mathbf{P}(\Omega) = 1$ and that $\mathbf{P}(\phi) = 0$.

It is also clear from our previous discussions that the whole point of probability is to say how likely the various outcomes are, either individually or, more usually, collectively in events. Here are some more examples to illustrate this.

Example: Opinion Poll Suppose n people are picked at random and interrogated as to their opinion (like or dislike or do not care) about a brand of toothpaste. Here, the sample space Ω is all collections of three integers (x, y, z) such that $x + y + z = n$, where x is the number that like it, y the number that dislike it, and z the number that do not care. Here an event of interest is

$$A \equiv \text{more like it than dislike it}$$

which comprises all the triples (x, y, z) with $x > y$. Another event that may worry the manufacturers is

$$B \equiv \text{the majority of people do not care,}$$

which comprises all triples (x, y, z) , such that $x + y < z$. ●

Example: Picking a Lottery Number In one conventional lottery, entries and the draw choose six numbers from 49. The sample space is therefore all sextuples $\{x_1, \dots, x_6\}$, where all the entries are between 1 and 49, and no two are equal. The principal event of interest is that this is the same as your choice. ●

Example: Coins If a coin is tossed once, then $\Omega = \{\text{head, tail}\}$. In line with the notation above, we usually write

$$\Omega = \{H, T\}.$$

The event that the coin shows a head should strictly be denoted by $\{H\}$, but in common with most other writers we omit the braces in this case, and denote a head by H . Obviously, $H^c = T$ and $T^c = H$.

Likewise, if a coin is tossed twice, then

$$\Omega = \{HH, HT, TH, TT\},$$

and so on. This experiment is performed even more often in probability textbooks than it is in real life. ●

Because events are sets, we use the usual notation for combining them; thus:

$A \cap B$ denotes outcomes in both A and B ; their *intersection*.

$A \cup B$ denotes outcomes in either A or B or both; their *union*.

$A \Delta B$ denotes outcomes in either A or B , but not both; their *symmetric difference*.

$A \setminus B$ denotes outcomes in A that are not in B ; their *difference*.

$\bigcup_{j=1}^{\infty} A_j$ denotes outcomes that are in at least one of the countable collection $(A_j; j \geq 1)$; their *countable union*.

[Countable sets are in one–one correspondence with a subset of the positive integers.]

$A \subseteq B$ denotes that every outcome in A is also in B ; this is *inclusion*.

$A = \{\omega_1, \omega_2, \omega_3, \dots, \omega_n\}$ denotes that the event A consists of the outcomes $\omega_1, \dots, \omega_n$

$A \times B$ denotes the product of A and B ; that is, the set of all ordered pairs (ω_a, ω_b) , where $\omega_a \in A$ and $\omega_b \in B$.

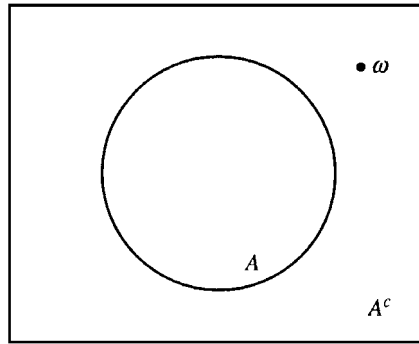


Figure 1.1 The interior of the rectangle represents the sample space Ω , and the interior of the circle represents an event A . The point ω represents an outcome in the event A^c . The diagram clearly illustrates the identities $A^c \cup A = \Omega$ and $\Omega \setminus A = A^c$.

These methods of combining events give rise to many equivalent ways of denoting an event. Some of the more useful identities for any events A and B are:

- (1) $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$
- (2) $A = (A \cap B) \cup (A \cap B^c)$
- (3) $A \setminus B = A \cap B^c$
- (4) $A^c = \Omega \setminus A$
- (5) $A \cap A^c = \phi$
- (6) $A \cup A^c = \Omega$.

These identities are easily verified by checking that every element of the left-hand side is included in the right-hand side, and vice versa. You should do this.

Such relationships are often conveniently represented by simple diagrams. We illustrate this by providing some basic examples in Figures 1.1 and 1.2. Similar relationships hold

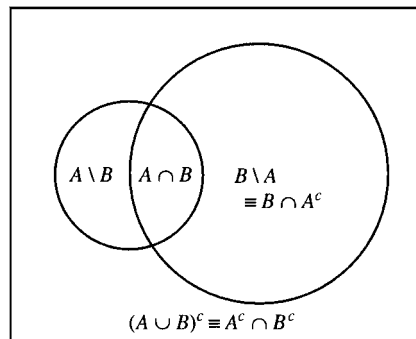


Figure 1.2 The interior of the smaller circle represents the event A ; the interior of the larger circle represents the event B . The diagram illustrates numerous simple relationships; for example, the region common to both circles is $A \cap B \equiv (A^c \cup B^c)^c$. For another example, observe that $A \Delta B = A^c \Delta B^c$.

between combinations of three or more events and some of these are given in the problems at the end of this chapter.

When $A \cap B = \phi$ we say that A and B are disjoint (or mutually exclusive).

- (7) **Example** A die is rolled. The outcome is one of the integers from 1 to 6. We may denote these by $\{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$, or more directly by $\{1, 2, 3, 4, 5, 6\}$, as we choose. Define:

A the event that the outcome is even,

B the event that the outcome is odd,

C the event that the outcome is prime,

D the event that the outcome is perfect (a perfect number is the sum of its prime factors).

Then the above notation compactly expresses obvious statements about these events. For example:

$$\begin{aligned} A \cap B &= \phi & A \cup B &= \Omega \\ A \cap D &= \{\omega_6\} & C \setminus A &= B \setminus \{\omega_1\} \end{aligned}$$

and so on. ●

It is natural and often useful to consider the number of outcomes in an event A . This is denoted by $|A|$, and is called the *size* or *cardinality* of A .

It is straightforward to see, by counting the elements on each side, that size has the following properties.

If A and B are disjoint, then

$$(8) \quad |A \cup B| = |A| + |B|,$$

and more generally, for any A and B

$$(9) \quad |A \cup B| + |A \cap B| = |A| + |B|.$$

If $A \subseteq B$, then

$$(10) \quad |A| \leq |B|.$$

For the product $A \times B$,

$$(11) \quad |A \times B| = |A||B|.$$

Finally,

$$(12) \quad |\phi| = 0.$$

- (13) **Example** The Shelmikedmu are an elusive and nomadic tribe whose members are unusually heterogeneous in respect of hair and eye colour, and skull shape. A persistent anthropologist establishes the following facts:

- (i) 75% have dark hair, the rest have fair hair.
- (ii) 80% have brown eyes; the rest have blue eyes.

- (iii) No narrow-headed person has fair hair and blue eyes.
- (iv) The proportion of blue-eyed broad-headed tribespeople is the same as the proportion of blue-eyed narrow-headed tribespeople.
- (v) Those who are blue-eyed and broad-headed are fair-haired or dark-haired in equal proportion.
- (vi) Half the tribe is dark-haired and broad-headed.
- (vii) The proportion who are brown-eyed, fair-haired, and broad-headed is equal to the proportion who are brown eyed, dark-haired, and narrow-headed.

The anthropologist also finds n , the proportion of the tribe who are narrow-headed, but unfortunately this information is lost in a clash with a crocodile on the difficult journey home. Is another research grant and field trip required to find n ? Fortunately, not if the anthropologist uses set theory. Let

B be the set of those with blue eyes

C be the set of those with narrow heads

D be the set of those with dark hair

Then the division of the tribe into its heterogeneous sets can be represented by Figure 1.3. This type of representation of sets and their relationships is known as a *Venn diagram*. The proportion of the population in each set is denoted by the lower case letter in each compartment, so

$$a = |B^c \cap C^c \cap D^c|/|\Omega|,$$

$$b = |B \cap C^c \cap D^c|/|\Omega|,$$

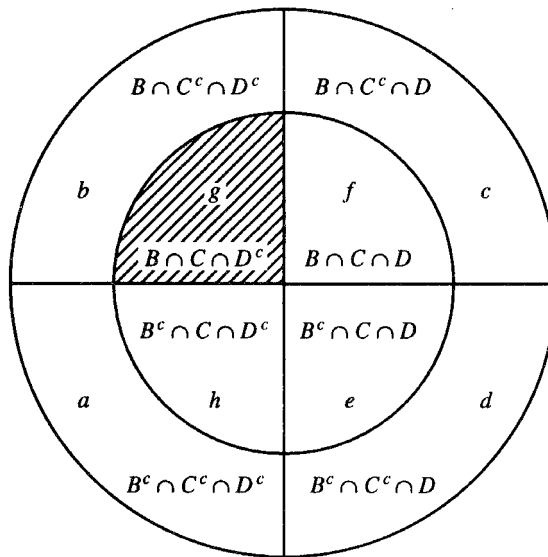


Figure 1.3 Here the interior of the large circle represents the entire tribe, and the interior of the small circle represents those with narrow heads. The part to the right of the vertical line represents those with dark hair, and the part above the horizontal line represents those with blue eyes. Thus, the shaded quadrant represents those with blue eyes, narrow heads, and fair hair; as it happens, this set is empty by (iii). That is to say $B \cap C \cap D^c = \phi$, and so $g = 0$.

and so on. The required proportion having narrow heads is

$$n = |C|/|\Omega| = e + f + g + h$$

and, of course, $a + b + c + d + e + f + g + h = 1$. The information in (i)–(vii), which survived the crocodile, yields the following relationships:

- | | |
|-------|------------------------|
| (i) | $c + d + e + f = 0.75$ |
| (ii) | $a + d + e + h = 0.8$ |
| (iii) | $g = 0$ |
| (iv) | $f + g = b + c$ |
| (v) | $b = c$ |
| (vi) | $c + d = 0.5$ |
| (vii) | $a = e$ |

The anthropologist (who has a pretty competent knowledge of algebra) solves this set of equations to find that

$$n = e + f + g + h = e + f + h = 0.15 + 0.1 + 0.05 = 0.3$$

Thus, three-tenths of the tribe are narrow-headed. ●

This section concludes with a technical note (which you may omit on a first reading). We have noted that events are subsets of Ω . A natural question is, which subsets of Ω are entitled to be called events?

It seems obvious that if A and B are events, then $A \cup B$, A^c , $A \cap B$, and so on should also be entitled to be events. This is a bit vague; to be precise, we say that a subset A of Ω can be an event if it belongs to a collection \mathcal{F} of subsets of Ω , obeying the following three rules:

- | | |
|------|----------------------------------------------------------------------------------------------|
| (14) | $\Omega \in \mathcal{F};$ |
| (15) | if $A \in \mathcal{F}$ then $A^c \in \mathcal{F};$ |
| (16) | if $A_j \in \mathcal{F}$ for $j \geq 1$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}.$ |

The collection \mathcal{F} is called an event space or a σ -field.

Notice that using (1)–(6) shows that if A and B are in \mathcal{F} , then so are $A \setminus B$, $A \Delta B$ and $A \cap B$.

- (17) **Example (7) Revisited** It is easy for you to check that $\{\phi, A, B, \Omega\}$ is an event space, and $\{\phi, A \cup C, B \setminus C, \Omega\}$ is an event space. However, $\{\phi, A, \Omega\}$ and $\{\phi, A, B, D, \Omega\}$ are *not* event spaces. ●

In general, if Ω is finite, it is quite usual to take \mathcal{F} to be the collection of all subsets of Ω , which is clearly an event space. If Ω is infinite, then this collection is sometimes too big to be useful, and some smaller collection of subsets is required.

1.3 The Addition Rules for Probability

An event A has probability $\mathbf{P}(A)$. But any experiment may give rise to a great many events of interest, and we have seen in Section 1.2 that these can be combined in numerous ways. We need some rules that tell us how to deal with probabilities of complicated events. Naturally, we continue to require that for any event A

$$(1) \quad 0 \leq \mathbf{P}(A) \leq 1$$

and, in particular, that the certain event has probability 1, so

$$(2) \quad \mathbf{P}(\Omega) = 1.$$

It turns out that we require only one type of rule, the addition rules. The simplest form of this is as follows:

The Basic Addition Rule If A and B are disjoint events, then

$$(3) \quad \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

This rule lies at the heart of probability. First, let us note that we need such a rule, because $A \cup B$ is an event when A and B are events, and we therefore need to know its probability. Second, note that it follows from (3) (by induction) that if A_1, A_2, \dots, A_n is any collection of disjoint events, then

$$(4) \quad \mathbf{P}\left(\bigcup_{i=1}^n A_i\right) = \mathbf{P}(A_1) + \dots + \mathbf{P}(A_n).$$

The proof is a simple exercise, using induction.

Third, note that it is sometimes too restrictive to confine ourselves to a finite collection of events (we have seen several sample spaces, with infinitely many outcomes), and we therefore need an extended version of (4).

Extended Addition Rule If A_1, A_2, \dots is a collection of disjoint events, then

$$(5) \quad \mathbf{P}(A_1 \cup A_2 \cup \dots) = \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots.$$

Equation (5), together with (1) and (2),

$$0 \leq \mathbf{P}(A) \leq 1 \quad \text{and} \quad \mathbf{P}(\Omega) = 1,$$

are sometimes said to be the *axioms* of probability. They describe the behaviour of the probability function \mathbf{P} defined on subsets of Ω . In fact, in everyday usage, \mathbf{P} is not referred to as a probability function but as a probability distribution. Formally, we state the following.

Definition Let Ω be a sample space, and suppose that $\mathbf{P}(\cdot)$ is a probability function on a family of subsets of Ω satisfying (1), (2), and (5). Then \mathbf{P} is called a probability distribution on Ω . ▲

The word distribution is used because it is natural to think of probability as something that is distributed over the outcomes in Ω . The function \mathbf{P} tells you just how it is distributed. In this respect, probability behaves like distributed mass, and indeed in many books authors do speak of a unit of probability mass being distributed over the sample space, and refer to \mathbf{P} as a probability mass function. This metaphor can be a useful aid to intuition because, of course, mass obeys exactly the same addition rule. If two distinct objects A and B have respective masses $m(A)$ and $m(B)$, then the mass of their union $m(A \cup B)$ satisfies

$$m(A \cup B) = m(A) + m(B).$$

Of course, mass is also nonnegative, which reinforces the analogy.

We conclude this section by showing how the addition rule is consistent with, and suggested by, our interpretations of probability as a proportion.

First, consider an experiment with equally likely outcomes, for which we defined probability as the proportion

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

If A and B are disjoint then, trivially, $|A \cup B| = |A| + |B|$. Hence, in this case,

$$\mathbf{P}(A \cup B) = \frac{|A \cup B|}{|\Omega|} = \frac{|A|}{|\Omega|} + \frac{|B|}{|\Omega|} = \mathbf{P}(A) + \mathbf{P}(B).$$

Second, consider the interpretation of probability as reflecting relative frequency in the long run. Suppose an experiment is repeated N times. At each repetition, events A and B may or may not occur. If they are disjoint, they cannot both occur at the same repetition. We argued in Section 0.4 that the relative frequency of any event should be not too far from its probability. Indeed, it is often the case that the relative frequency $N(A)/N$ of an event A is the only available guide to its probability $\mathbf{P}(A)$. Now, clearly

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

Hence, dividing by N , there is a powerful suggestion that we should have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

Third, consider probability as a measure of expected value. For this case, we resurrect the benevolent plutocrat who is determined to give away \$1 at random. The events A and B are disjoint. If A occurs, you get \$1 in your left hand; if B occurs you get \$1 in your right hand. If $(A \cup B)^c$ occurs, then Jack gets \$1. The value of this offer to you is $\mathbf{P}(A \cup B)$; the value to your left hand is $\mathbf{P}(A)$; and the value to your right hand is $\mathbf{P}(B)$. Obviously, it does not matter in which hand you get the money, so

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

Finally, consider the case where we imagine a point is picked at random anywhere in some plane region Ω of area $|\Omega|$. If $A \subseteq \Omega$, we defined

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

Because area also satisfies the addition rule, we have immediately, when $A \cap B = \phi$, that

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

It is interesting and important to note that in this case the analogy with mass requires the unit probability mass to be distributed uniformly over the region Ω . We can envisage this distribution as a lamina of uniform density $|\Omega|^{-1}$ having total mass unity. This may seem a bizarre thing to imagine, but it turns out to be useful later.

In conclusion, it seems that the addition rule is natural and compelling in every case where we have any insight into the behaviour of probability. Of course, it is a big step to say that it should apply to probability in every other case, but it seems inevitable. Doing so has led to remarkably elegant and accurate descriptions of the real world.

This property (4) is known as *finite additivity* and (5) is *countable additivity*. Note that if $A \subseteq B$, then

$$(6) \quad \mathbf{P}(A) \leq \mathbf{P}(B)$$

and finally, using $A \cup A^c = \Omega$, we have

$$(7) \quad \mathbf{P}(\phi) = 0.$$

Once again, these statements are quite consistent with our intuition about likelihoods, as reinforced by experience.

Historically, the theory of probability has its roots firmly based in observation of games of chance employing cards, dice, and lotteries.

- (8) **Example** Three dice are rolled and the numbers on the upper faces are added together. The outcomes 9 and 10 can each be obtained in six distinct ways; thus:

$$\begin{aligned} 10 &= 1 + 3 + 6 = 1 + 4 + 5 = 2 + 2 + 6 = 2 + 3 + 5 = 2 + 4 + 4 = 3 + 3 + 4 \\ 9 &= 1 + 2 + 6 = 1 + 3 + 5 = 1 + 4 + 4 = 2 + 2 + 5 = 2 + 3 + 4 = 3 + 3 + 3. \end{aligned}$$

Some time before 1642, Galileo was asked to explain why, despite this, the outcome 10 is more likely than the outcome 9, as shown by repeated experiment. He observed that the sample space Ω has $6^3 = 216$ outcomes, being all possible triples of numbers from 1 to 6. Of these, 27 sum to 10, and 25 sum to 9, so $\mathbf{P}(10) = \frac{27}{216}$ and $\mathbf{P}(9) = \frac{25}{216}$.

This provides an explanation for the preponderance of 10 over 9. ●

It is just this kind of agreement between theory and experiment that justifies our adoption of the rules above. We will see many more examples of this.

1.4 Properties of Probability

We have agreed that, given a space \mathcal{F} of events A , the probability function $\mathbf{P}(\cdot)$ satisfies the following rules (or axioms) that we display as a definition.

Definition The function $\mathbf{P}(\cdot) : \mathcal{F} \rightarrow [0, 1]$ is a probability function if

$$(1) \quad \mathbf{P}(A) \geq 0 \quad \text{for all } A \in \mathcal{F},$$

$$(2) \quad \mathbf{P}(\Omega) = 1$$

and

$$(3) \quad \mathbf{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbf{P}(A_j)$$

whenever A_1, A_2, \dots are disjoint events (which is to say that $A_i \cap A_j = \phi$ whenever $i \neq j$). ▲

In passing, we note that (3) is known as the property of *countable additivity*. Obviously, it implies *finite additivity* so that, in particular, if $A \cap B = \phi$, then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

From these three rules we can derive many important and useful relationships, for example,

$$(4) \quad \mathbf{P}(\phi) = 0$$

$$(5) \quad \mathbf{P}(A^c) = 1 - \mathbf{P}(A),$$

$$(6) \quad \mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(A \cap B),$$

$$(7) \quad \mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B),$$

$$(8) \quad \mathbf{P}\left(\bigcup_1^n A_i\right) = \sum_1^n \mathbf{P}(A_i) - \sum_{i < j} \mathbf{P}(A_i \cap A_j) + \sum_{i < j < k} \mathbf{P}(A_i \cap A_j \cap A_k) + \dots \\ + (-)^{n+1} \mathbf{P}(A_1 \cap A_2 \cap \dots \cap A_n).$$

The following examples begin to demonstrate the importance and utility of (1)–(8).

(9) **Example** Let us prove (4), (5), and (7) above. First, by (2) and (3), for any $A \in \mathcal{F}$,

$$1 = \mathbf{P}(\Omega) = \mathbf{P}(A \cup A^c) = \mathbf{P}(A) + \mathbf{P}(A^c)$$

which proves (5). Now, setting $A = \Omega$ establishes (4). Finally, using (3) repeatedly, we obtain

$$(10) \quad \mathbf{P}(B) = \mathbf{P}(B \cap A) + \mathbf{P}(B \cap A^c),$$

and

$$\mathbf{P}(A \cup B) = \mathbf{P}(A \cup (B \cap A^c)) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(B \cap A) \\ \text{by (1),}$$

which proves (7). ●

You should now prove (6) as an elementary exercise; the proof of (8) is part of Problem 12.

(11) **Example: Inequalities for $\mathbf{P}(\cdot)$** (i) If $A \subseteq B$ then $B \cap A = A$, so from (10)

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \cap A^c) \geq \mathbf{P}(A) \quad \text{by (1).}$$

(ii) For any A, B , we have *Boole's inequalities*

$$\begin{aligned} \mathbf{P}(A) + \mathbf{P}(B) &\geq \mathbf{P}(A \cup B) && \text{by (7).} \\ &\geq \max\{\mathbf{P}(A), \mathbf{P}(B)\} && \text{by part (i)} \\ &\geq \mathbf{P}(A \cap B) && \text{by part (i) again} \\ &\geq \mathbf{P}(A) + \mathbf{P}(B) - 1 && \text{by (7) again} \end{aligned} \quad \bullet$$

(13) **Example: Lottery** An urn contains 1000 lottery tickets numbered from 1 to 1000. One is selected at random. A fairground performer offers to pay \$3 to anyone who has already paid him \$2, if the number on the ticket is divisible by 2, 3, or 5. Would you pay him your \$2 before the draw? (If the ticket number is not divisible by 2, 3, or 5 you lose your \$2.)

Solution Let D_k be the event that the number drawn is divisible by k . Then

$$\mathbf{P}(D_2) = \frac{500}{1000} = \frac{1}{2}$$

and so on. Also

$$\mathbf{P}(D_2 \cap D_3) = \mathbf{P}(D_6) = \frac{166}{1000}$$

and so forth. Using (8) with $n = 3$ and making several similar calculations, we have

$$\begin{aligned} \mathbf{P}(D_2 \cup D_3 \cup D_5) &= \mathbf{P}(D_2) + \mathbf{P}(D_3) + \mathbf{P}(D_5) + \mathbf{P}(D_2 \cap D_3 \cap D_5) \\ &\quad - \mathbf{P}(D_2 \cap D_3) - \mathbf{P}(D_3 \cap D_5) - \mathbf{P}(D_2 \cap D_5) \\ &= 10^{-3}(500 + 333 + 200 + 33 - 166 - 66 - 100) = \frac{367}{500}. \end{aligned}$$

The odds on winning are thus better than 2:1, and you should accept his generous offer. ●

1.5 Sequences of Events

This section is important, but may be omitted at a first reading.

Often, we are confronted by an infinite sequence of events $(A_n; n \geq 1)$ such that $A = \lim_{n \rightarrow \infty} A_n$ exists. In particular, if $A_n \subseteq A_{n+1}$ for all n , then

$$(1) \quad \lim_{n \rightarrow \infty} A_n = \bigcup_{j=1}^{\infty} A_j = A,$$

and A is an event by (1.2.16). It is of interest, and also often useful, to know $\mathbf{P}(A)$. The following theorem is therefore important as well as attractive.

(2) **Theorem** If $A_n \subseteq A_{n+1} \in \mathcal{F}$ for all $n \geq 1$ and $A = \lim_{n \rightarrow \infty} A_n$, then

$$\mathbf{P}(A) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n).$$

Proof. Because $A_n \subseteq A_{n+1}$, we have $(A_{j+1} \setminus A_j) \cap (A_{k+1} \setminus A_k) = \phi$, for $k \neq j$. Also, setting $A_0 = \phi$

$$\bigcup_1^n (A_j \setminus A_{j-1}) = A_n.$$

Furthermore,

$$(3) \quad \mathbf{P}(A_{n+1} \setminus A_n) = \mathbf{P}(A_{n+1}) - \mathbf{P}(A_n).$$

Hence, because $A_0 = \phi$

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}\left(\lim_{n \rightarrow \infty} A_n\right) = \mathbf{P}\left(\bigcup_{j=0}^{\infty} A_{j+1} \setminus A_j\right) \\ &= \sum_{j=0}^{\infty} \mathbf{P}(A_{j+1} \setminus A_j) \quad \text{by (1.4.3)} \\ &= \lim_{n \rightarrow \infty} \mathbf{P}(A_n) \quad \text{by (3).} \quad \blacksquare \end{aligned}$$

From this result, it is a simple matter to deduce that if $A_n \supseteq A_{n+1}$ for all n , then

$$(4) \quad \lim_{n \rightarrow \infty} \mathbf{P}(A_n) = \mathbf{P}(A).$$

With a bit more work, one can show more generally that if $\lim_{n \rightarrow \infty} A_n = A$, then (4) is still true. Because of this, the function $\mathbf{P}(\cdot)$ is said to be a *continuous set function*.

1.6 Remarks

Simple problems in probability typically require the calculation of the probability $\mathbf{P}(E)$ of some event E , or at least the calculation of bounds for $\mathbf{P}(E)$. The underlying experiment may be implicit or explicit; in any case, the first step is always to choose the sample space. (Naturally, we choose the one that makes finding $\mathbf{P}(E)$ easiest.)

Then you may set about finding $\mathbf{P}(E)$ by using the rules and results of Section 1.4, and the usual methods for manipulating sets. Useful aids at this simple level include Venn diagrams, and identities such as de Morgan's laws, namely:

$$(1) \quad \left(\bigcup_{i=1}^n A_i\right)^c = \bigcap_{i=1}^n A_i^c$$

and

$$(2) \quad \left(\bigcap_{i=1}^n A_i\right)^c = \bigcup_{i=1}^n A_i^c.$$

These are readily established directly or by induction, as follows. First, draw a Venn diagram to see that $(A \cup B)^c = A^c \cap B^c$.

Now if $\bigcup_{j=1}^n A_j = B_n$, then

$$\left(\bigcup_{j=1}^{n+1} A_j \right)^c = (A_{n+1} \cup B_n)^c = A_{n+1}^c \cap B_n^c.$$

Hence, (1) follows by induction on n . The result (2) can also be proved directly, by induction, or by using (1). You can do this, and you should do it now.

We end this section with a note for the more demanding reader. It has been claimed above that probability theory is of wide applicability, yet most of the examples in this chapter deal with the behaviour of coins, dice, lotteries, and urns. For most of us, weeks or even months can pass without any involvement with dice or urns. (The author has never even seen an urn, let alone removed a red ball from one.) The point is that such simple problems present their probabilistic features to the reader unencumbered by strained assumptions about implausible models of reality. The penalty for simplicity is that popular problems may become hackneyed through overuse in textbooks. We take this risk, but reassure the reader that more realistic problems feature largely in later chapters. In addition, when considering some die or urn, students may be pleased to know that they are treading in the footsteps of many eminent mathematicians as they perform these calculations. Euler or Laplace may have pondered over exactly the same difficulty as you, although perhaps not for so long.

1.7 Review and Checklist for Chapter 1

We began by introducing the idea of experiments and their outcomes, with events and their probabilities. Then we used our experience of chance events, and intuitive ideas about probability as a proportion, to formulate a set of rules (or axioms) for probability. Of these, the addition rule is the most important, and we derived several useful corollaries that will be in constant use. In the course of this, we also introduced some standard notation. For convenience, the key rules and results are summarized here.

Notation

Ω	sample space of outcomes \equiv certain event
A, B, C	possible events in Ω
ϕ	impossible event
$\mathbf{P}(\cdot)$	the probability function
$\mathbf{P}(A)$	the probability that A occurs
$A \cup B$	either A or B occurs, or both occur (union)
$A \cap B$	both A and B occur (intersection)
A^c	A does not occur (complement)
$A \setminus B$	difference: A occurs and B does not

$A \Delta B$	either A or B , but not both
$A \subseteq B$	if A occurs, then B occurs
$\{A_i\}$	is a partition of Ω if the A_i are disjoint with union Ω
$A \cap B = \phi$	A and B are disjoint
\mathcal{F}	event space (or σ -field)
$ A $	cardinality (or size) of A

Axioms and Rules

Any subset A of Ω is an event if it belongs to a collection \mathcal{F} of subsets of Ω such that

- $\Omega \in \mathcal{F}$
- if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
- if $A_i \in \mathcal{F}$ for $i \geq 1$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A function $\mathbf{P}(\cdot)$ defined on events is a probability function or probability distribution if

- $\mathbf{P}(A) \geq 0$ for all A in \mathcal{F}
- $\mathbf{P}(\Omega) = 1$
- $\mathbf{P}\left(\bigcup_1^{\infty} A_i\right) = \sum_i \mathbf{P}(A_i)$ for any disjoint events $\{A_i\}$.

The function $\mathbf{P}(\cdot)$ obeys the following rules:

Range: $0 \leq \mathbf{P}(A) \leq 1$

Impossible event: $\mathbf{P}(\phi) = 0$

Inclusion-exclusion: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$

Complement: $\mathbf{P}(A) = 1 - \mathbf{P}(A^c)$

Difference: $\mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(B)$ when $B \subseteq A$

General inclusion-exclusion:

$$\mathbf{P}\left(\bigcup_1^n A_i\right) = \sum_i \mathbf{P}(A_i) - \sum_{i < j} \mathbf{P}(A_i \cap A_j) + \cdots - (-1)^n \mathbf{P}\left(\bigcap_1^n A_i\right).$$

Boole's inequalities:

$$\mathbf{P}\left(\bigcup_1^n A_i\right) \leq \sum_i \mathbf{P}(A_i).$$

Checklist of Terms for Chapter 1

- 1.1 experiment
 - outcome
 - sample space
 - event
 - probability
 - Venn diagram
- 1.2 certain event
 - impossible event

event space

σ -field

1.3 addition rules

probability distribution

countable additivity

1.4 axioms of probability

Boole's inequalities

1.5 continuous set function

1.6 de Morgan's Laws

WORKED EXAMPLES AND EXERCISES

1.8 Example: Dice

You roll two dice. What is the probability of the events:

- (a) They show the same?
- (b) Their sum is seven or eleven?
- (c) They have no common factor greater than unity?

Solution First, we must choose the sample space. A natural representation is as ordered pairs of numbers (i, j) , where each number refers to the face shown by one of the dice. We require the dice to be distinguishable (one red and one green, say) so that $1 \leq i \leq 6$ and $1 \leq j \leq 6$. Because of the symmetry of a perfect die, we assume that these 36 outcomes are equally likely.

- (a) Of these 36 outcomes, just six are of the form (i, i) , so using (1.3.1) the required probability is

$$\frac{|A|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

- (b) There are six outcomes of the form $(i, 7 - i)$ whose sum is 7, so the probability that the sum is 7 is $\frac{6}{36} = \frac{1}{6}$.

There are two outcomes whose sum is 11—namely, $(5, 6)$ and $(6, 5)$ —so the probability that the sum is 11 is $\frac{2}{36} = \frac{1}{18}$.

Hence, using (1.4.3), the required probability is $\frac{1}{6} + \frac{1}{18} = \frac{2}{9}$.

- (c) It is routine to list the outcomes that do have a common factor greater than unity. They are 13 in number, namely:

$$\{(i, i); i \geq 2\}, (2, 4), (4, 2), (2, 6), (6, 2), (3, 6), (6, 3), (4, 6), (6, 4).$$

This is the complementary event, so by (1.4.5) the required probability is

$$1 - \frac{13}{36} = \frac{23}{36}.$$

Doing it this way gives a slightly quicker enumeration than the direct approach.

- (1) **Exercise** What is the probability that the sum of the numbers is 2, 3, or 12?
- (2) **Exercise** What is the probability that:
 - (a) The sum is odd?
 - (b) The difference is odd?
 - (c) The product is odd?
- (3) **Exercise** What is the probability that one number divides the other?
- (4) **Exercise** What is the probability that the first die shows a smaller number than the second?
- (5) **Exercise** What is the probability that different numbers are shown and the smaller of the two numbers is r , $1 \leq r \leq 6$?

Remark It was important to distinguish the two dice. Had we not done so, the sample space would have been $\{(i, j); 1 \leq i \leq j \leq 6\}$, and these 21 outcomes are not equally likely, either intuitively or empirically. Note that the dice need not be different colours in fact, it is enough for us to be able to suppose that they are.

1.9 Example: Urn

An urn contains n heliotrope and n tangerine balls. Two balls are removed from the urn together, at random.

- (a) What is the sample space?
- (b) What is the probability of drawing two balls of different colours?
- (c) Find the probability p_n that the balls are the same colour, and evaluate $\lim_{n \rightarrow \infty} p_n$.

Solution (a) As balls of the same colour are otherwise indistinguishable, and they are drawn together, a natural sample space is $\Omega = \{HH, HT, TT\}$.

(b) The outcomes in Ω exhibit no symmetry. Taking our cue from the previous example, we choose to distinguish the balls by numbering them from 1 to $2n$, and also suppose that they are drawn successively. Then the sample space is the collection of ordered pairs of the form (i, j) , where $1 \leq i, j \leq 2n$, and $i \neq j$, because we cannot pick the same ball twice. These $2n(2n - 1)$ outcomes are equally likely, by symmetry. In n^2 of them, we draw heliotrope followed by tangerine, and in n^2 , we draw tangerine followed by heliotrope. Hence,

$$\mathbf{P}(HT) = \frac{n^2 + n^2}{2n(2n - 1)} = \frac{n}{2n - 1}.$$

(c) By (1.4.5)

$$p_n = 1 - \mathbf{P}(HT) = \frac{n - 1}{2n - 1} \rightarrow \frac{1}{2} \text{ as } n \rightarrow \infty.$$

- (1) **Exercise** Find $\mathbf{P}(HH)$ when the sample space is taken to be all unordered pairs of distinguishable balls.
- (2) **Exercise** What is the probability that (a) the first ball is tangerine? (b) the second ball is tangerine?
- (3) **Exercise** Half the balls are removed and placed in a box. One of those remaining in the urn is removed. What is the probability that it is tangerine?

- (4) **Exercise** A fair die with n sides is rolled. If the r th face is shown, r balls are removed from the urn and placed in a bag. What is the probability that a ball removed at random from the bag is tangerine?

1.10 Example: Cups and Saucers

A tea set has four cups and saucers with two cups and saucers in each of two different colours. If the cups are placed at random on the saucers, what is the probability that no cup is on a saucer of the same colour?

Solution I Call the colours azure and blue; let A be the event that an azure cup is on an azure saucer, and B the event that a blue cup is on a blue saucer. Because there are only two places for blue cups not to be on blue saucers, we see that A occurs if and only if B occurs, so $A = B$ and

$$\mathbf{P}((A \cup B)^c) = \mathbf{P}(A^c) = 1 - \mathbf{P}(A).$$

Now $\mathbf{P}(A) = \mathbf{P}(A_1 \cup A_2)$, where A_1 and A_2 denote the events that the first and second azure cups, respectively, are on azure saucers. There are 24 equally probable ways of putting the four cups on the four saucers. In 12 of them, A_1 occurs; in 12 of them, A_2 occurs; and in 4 of them, $A_1 \cap A_2$ occurs, by enumeration.

Hence, by (1.4.7),

$$\mathbf{P}(A) = \frac{12}{24} + \frac{12}{24} - \frac{4}{24} = \frac{5}{6},$$

and the required probability is $\frac{1}{6}$.

Solution II Alternatively, instead of considering all the ways of placing cups on saucers, we may consider only the distinct ways of arranging the cups by colour with the saucers fixed. There are only six of these, namely:

$$aabb; abba; abab; baab; baba; bbaa;$$

and by symmetry they are equally likely. By inspection, in only one of these arrangements is no cup on a saucer of the same colour, so the required probability is $\frac{1}{6}$.

Remark In this example, considering a *smaller* sample space makes the problem easier. This is in contrast to our solutions to Examples 1.7 and 1.8, where we used larger sample spaces to simplify things.

- (1) **Exercise** What is the probability that exactly
- One cup is on a saucer of the same colour?
 - Two cups are on saucers of the same colour?
- (2) **Exercise** What is the probability that no cup is on a saucer of the same colour if the set comprises four cups and saucers in four distinct colours?

1.11 Example: Sixes

Three players, Achilles, Briseis, and Chryseis, take turns to roll a die in the order ABC, ABC, A . . . Each player drops out of the game immediately upon throwing a six.

- (a) What is the sample space for this experiment?
- (b) Suppose the game stops when two players have rolled a six. What is the sample space for this experiment?
- (c) What is the probability that Achilles is the second player to roll a six?
- (d) Let D_n be the event that the third player to roll a six does so on the n th roll. Describe the event E given by

$$E = \left(\bigcup_{n=1}^{\infty} D_n \right)^c .$$

Solution (a) Let U be the collection of all sequences x_1, \dots, x_n for all $n \geq 1$, such that

$$\begin{aligned} x_j &\in \{1, 2, 3, 4, 5\} \quad \text{for } 1 \leq j < n, \\ x_n &= 6. \end{aligned}$$

Then each player's rolls generate such a sequence, and the sample space consists of all selections of the triple (u_1, u_2, u_3) , where $u_i \in U$ for $1 \leq i \leq 3$. We may denote this by $U \times U \times U$ or even U^3 , if we want.

(b) Let V be the collection of all sequences x_1, x_2, \dots, x_n , for $n \geq 1$, such that

$$x_j \in \{1, 2, 3, 4, 5\} \quad \text{for } 1 \leq j \leq n.$$

Then the sample space consists of two selections u_1, u_2 from U , corresponding to the players who roll sixes, and one selection v from V corresponding to the player who does not. The length of v equals the longer of u_1 and u_2 if this turn in the round comes before the second player to get a six, or it is one less than the longer of u_1 and u_2 if this turn in the round is later than the second player to get a six.

(c) Despite the answers to (a) and (b), we use a different sample space to answer this question. Suppose the player who is first to roll a six continues to roll the die when the turn comes round, these rolls being ignored by the others. This does not affect the respective chances of the other two to be the next player (of these two) to roll a six. We therefore let Ω be the sample space consisting of all sequences of length $3r + 1$, for $r \geq 0$, using the integers 1, 2, 3, 4, 5, or 6. This represents $3r + 1$ rolls of the die, and by the assumed symmetry the 6^{3r+1} possible outcomes are all equally likely for each r .

Suppose Achilles is the second player to roll a six on the $3r + 1$ th roll. Then his $r + 1$ rolls include no six except his last roll; this can occur in 5^r ways. If Briseis was first to roll a six, then her r rolls include at least one six; this may be accomplished in $6^r - 5^r$ ways. In this case, Chryseis rolled no six in r attempts; this can be done in 5^r ways. Hence, Achilles is second to Briseis in $5^r \cdot 5^r \cdot (6^r - 5^r)$ outcomes. Likewise, he is second to Chryseis in $5^r \cdot 5^r \cdot (6^r - 5^r)$ outcomes. Hence, the probability that he is second to roll a six on the $3r + 1$ th roll is

$$p_r = \frac{2(6^r - 5^r)5^{2r}}{6^{3r+1}}, \quad \text{for } r \geq 1.$$

By (1.4.3), therefore, the total probability that Achilles is second to roll a six is the sum of these, namely

$$\sum_{r=1}^{\infty} p_r = \frac{300}{1001}.$$

(d) The event $\bigcup_{n=1}^{\infty} D_n$ is the event that the game stops at the n th roll for some $n \geq 1$. Therefore, E is the event that they never stop.

- (1) **Exercise** For each player, find the probability that he or she is the first to roll a six.
- (2) **Exercise** Show that $\mathbf{P}(E) = 0$.
- (3) **Exercise** Find the probability that the Achilles rolls a six before Briseis rolls a six. (**Hint:** use a *smaller* sample space.)
- (4) **Exercise** Show that the probability that Achilles is last to throw a six is $\frac{305}{1001}$. Are you surprised that he is more likely to be last than to be second?

Remark The interesting thing about the solution to (c) is that the sample space Ω includes outcomes that are not in the original experiment, whose sample space is described in (a). The point is that the event in question has the same probability in the original experiment and in the modified experiment, but the required probability is obtained rather more easily in the second case because the outcomes are equally likely. This idea of augmenting the sample space was first used by Pascal and Fermat in the seventeenth century.

In fact, we find easier methods for evaluating this probability in Chapter 2, using new concepts.

1.12 Example: Family Planning

A woman planning her family considers the following schemes on the assumption that boys and girls are equally likely at each delivery:

- (a) Have three children.
- (b) Bear children until the first girl is born or until three are born, whichever is sooner, and then stop.
- (c) Bear children until there is one of each sex or until there are three, whichever is sooner, and then stop.

Let B_i denote the event that i boys are born, and let C denote the event that more girls are born than boys. Find $\mathbf{P}(B_1)$ and $\mathbf{P}(C)$ in each of the cases (a) and (b).

Solution (a) If we do not consider order, there are four distinct possible families: BBB , GGG , GGB , and BBG , but these are not equally likely. With order included, there are eight possible families in this larger sample space:

$$(1) \quad \{BBB; BBG; BGB; GBB; GGB; GBG; BGG; GGG\} = \Omega$$

and by symmetry they are equally likely. Now, by (1.3.1), $\mathbf{P}(B_1) = \frac{3}{8}$ and $\mathbf{P}(C) = \frac{1}{2}$. The fact that $\mathbf{P}(C) = \frac{1}{2}$ is also clear by symmetry.

Now consider (b). There are four possible families: $F_1 = G$, $F_2 = BG$, $F_3 = BBG$, and $F_4 = BBB$.

Once again, these outcomes are not equally likely, but as we have now done several times we can use a different sample space. One way is to use the sample space in (1), remembering that if we do this then some of the later births are fictitious. The advantage is that outcomes are equally likely by symmetry. With this choice, F_2 corresponds to $\{BGG \cup BGB\}$ and so $\mathbf{P}(B_1) = \mathbf{P}(F_2) = \frac{1}{4}$. Likewise, $F_1 = \{GGG \cup GGB \cup GBG \cup GBB\}$ and so $\mathbf{P}(C) = \frac{1}{2}$.

- (2) **Exercise** Find $\mathbf{P}(B_1)$ and $\mathbf{P}(C)$ in case (c).
- (3) **Exercise** Find $\mathbf{P}(B_2)$ and $\mathbf{P}(B_3)$ in all three cases.
- (4) **Exercise** Let E be the event that the completed family contains equal numbers of boys and girls. Find $\mathbf{P}(E)$ in all three cases.

1.13 Example: Craps

You roll two fair dice. If the sum of the numbers shown is 7 or 11, you win; if it is 2, 3, or 12, you lose. If it is any other number j , you continue to roll two dice until the sum is j or 7, whichever is sooner. If it is 7, you lose; if it is j , you win. What is the probability p that you win?

Solution Suppose that you roll the dice n times. That experiment is equivalent to rolling $2n$ fair dice, with the sample space Ω_{2n} being all possible sequences of length $2n$, of the numbers 1, 2, 3, 4, 5, 6, for any $n \geq 1$. By symmetry, these 6^{2n} outcomes are equally likely, and whether you win or you lose at or before the n th roll of the pair of dice is determined by looking at the sum of successive pairs of numbers in these outcomes.

The sample space for the roll of a pair of dice (Ω_2) has 36 equally likely outcomes. Let n_j denote the number of outcomes in which the sum of the numbers shown is j , $2 \leq j \leq 12$. Now let A_k be the event that you win by rolling a pair with sum k , and consider the 11 distinct cases:

(a)

$$\mathbf{P}(A_2) = \mathbf{P}(A_3) = \mathbf{P}(A_{12}) = 0,$$

because you always lose with these.

(b) For A_7 to occur, you must get 7 on the first roll. Because $n_7 = 6$

$$\mathbf{P}(A_7) = \frac{n_7}{|\Omega_2|} = \frac{6}{36} = \frac{1}{6}.$$

(c) Likewise,

$$\mathbf{P}(A_{11}) = \frac{n_{11}}{36} = \frac{2}{36} = \frac{1}{18}, \quad \text{since } n_{11} = 2.$$

(d) For A_4 to occur, you must get 4 on the first roll and on the n th roll, for some $n \geq 2$, with no 4 or 7 in the intervening $n - 2$ rolls. You can do this in $n_4^2 (36 - n_4 - n_7)^{n-2}$

ways and, therefore,

$$\begin{aligned} \mathbf{P}(A_4) &= \sum_{n=2}^{\infty} \frac{n_4^2(36 - n_4 - n_7)^{n-2}}{6^{2n}} && \text{by (1.4.3)} \\ &= \frac{n_4^2}{36(n_4 + n_7)} = \frac{1}{36} && \text{because } n_4 = 3. \end{aligned}$$

(e) Likewise,

$$\begin{aligned} \mathbf{P}(A_5) &= \frac{n_5^2}{36(n_5 + n_7)} \\ &= \frac{2}{45} && \text{because } n_5 = 4 \\ &= \mathbf{P}(A_9) && \text{because } n_9 = 4. \end{aligned}$$

Finally,

$$\begin{aligned} \mathbf{P}(A_6) &= \mathbf{P}(A_8) \\ &= \frac{25}{396} && \text{because } n_6 = n_8 = 5 \end{aligned}$$

and

$$\mathbf{P}(A_{10}) = \mathbf{P}(A_4) \quad \text{because } n_{10} = n_4 = 3.$$

Therefore, the probability that you win is

$$\begin{aligned} \mathbf{P}(A_7) + \mathbf{P}(A_{11}) + 2\mathbf{P}(A_4) + 2\mathbf{P}(A_5) + 2\mathbf{P}(A_6) &= \frac{1}{6} + \frac{1}{18} + \frac{1}{18} + \frac{4}{45} + \frac{25}{198} \\ &\simeq 0.493. \end{aligned}$$

- (1) **Exercise** What is the probability that you win on or before the second roll?
- (2) **Exercise** What is the probability that you win on or before the third roll?
- (3) **Exercise** What is the probability that you win if, on the first roll,
 - (a) The first die shows 2?
 - (b) The first die shows 6?
- (4) **Exercise** If you could fix the number to be shown by one die of the two on the first roll, what number would you choose?

1.14 Example: Murphy's Law

A fair coin is tossed repeatedly. Let s denote any fixed sequence of heads and tails of length r . Show that with probability one the sequence s will eventually appear in r consecutive tosses of the coin.

(The usual statement of Murphy's law says that anything that can go wrong, will go wrong).

Solution If a fair coin is tossed r times, there are 2^r distinct equally likely outcomes and one of them is s . We consider a fair die with 2^r faces; each face corresponds to one

of the 2^r outcomes of tossing the coin r times and one of them is face s . Now roll the die repeatedly.

Let A_k be the event that face s appears for the first time on the k th roll. There are 2^{rk} distinct outcomes of k rolls, and by symmetry they are equally likely. In $(2^r - 1)^{k-1}$ of them, A_k occurs, so by (1.3.1),

$$\mathbf{P}(A_k) = \frac{(2^r - 1)^{k-1}}{2^{rk}}.$$

Because $A_k \cap A_j = \phi$ for $k \neq j$, we have by (1.4.3) that

$$(1) \quad \mathbf{P}\left(\bigcup_1^m A_k\right) = \sum_1^m \mathbf{P}(A_k) = 1 - \left(\frac{2^r - 1}{2^r}\right)^m,$$

which is the probability that face s appears at all in m rolls.

Now consider n tosses of the coin, and let $m = [\frac{n}{r}]$ (where $[x]$ is the integer part of x). The n tosses can thus be divided into m sequences of length r with a remainder $n - mr$. Let B_n be the event that none of these m sequences is s , and let C_n be the event that the sequence s does not occur anywhere in the n tosses. Then

$$C_n \subseteq B_n = \left(\bigcup_1^m A_k\right)^c,$$

because rolling the die m times and tossing the coin mr times yield the same sample space of equally likely outcomes. Hence, by Example 1.4.11(i) and (1.4.5) and (1),

$$\mathbf{P}(C_n) \leq \mathbf{P}(B_n) = \left(\frac{2^r - 1}{2^r}\right)^m \rightarrow 0$$

as $n \rightarrow \infty$. Now the event that s eventually occurs is $\lim_{n \rightarrow \infty} (C_n^c)$, so by (1.4.5) and (1.5.4),

$$\mathbf{P}\left(\lim_{n \rightarrow \infty} C_n^c\right) = \lim_{n \rightarrow \infty} \mathbf{P}(C_n^c) = 1 - \lim_{n \rightarrow \infty} \mathbf{P}(C_n) = 1.$$

- (2) **Exercise** If the coin is tossed n times, show that the probability that it shows heads on an odd number of tosses (and tails on the rest) is $\frac{1}{2}$.
- (3) **Exercise** If the coin is tossed an unbounded number of times, show that the probability that a head is first shown on an odd numbered toss is $\frac{2}{3}$.
- (4) **Exercise** If Malone tosses his coin m times and Watt tosses his coin n times, show that the probability that they get the same number of heads each is equal to the probability that Beckett gets m heads in $m + n$ tosses of his coin.

PROBLEMS

N.B. Unless otherwise stated, coins are fair, dice are regular cubes and packs of cards are well shuffled with four suits of 13 cards.

- 1 You are given a conventional pack of cards. What is the probability that the top card is an ace?
- 2 You count a pack of cards (face down) and find it defective (having only 49 cards!). What is the probability that the top card is an ace?

- 3 A class contains seven boys and eight girls.
- If two are selected at random to leave the room, what is the probability that they are of different sexes?
 - On two separate occasions, a child is selected at random to leave the room. What is the probability that the two choices result in children of different sexes?
- 4 An urn contains 100 balls numbered from 1 to 100. Four are removed at random without being replaced. Find the probability that the number on the last ball is smaller than the number on the first ball.
- 5 Let \mathcal{F} be an event space. Show that the total number of events in \mathcal{F} cannot be exactly six. What integers can be the number of events in a finite event space?
- 6 To start playing a game of chance with a die, it is necessary first to throw a six.
- What is the probability that you throw your first six at your third attempt?
 - What is the probability that you require more than three attempts?
 - What is the most likely number of attempts until you first throw a six?
 - After how many throws would your probability of having thrown a six be at least 0.95?
- 7 Let A , B , and C be events. Write down expressions for the events where
- At least two of A , B , and C occur.
 - Exactly two of A , B , and C occur.
 - At most two of A , B , and C occur.
 - Exactly one of A , B , and C occurs.
- 8 A die is loaded in such a way that the probability that a 6 is thrown is five times that of any other number, each of them being equally probable.
- By what factor is the probability of a total of 24 from four throws greater than that for an unloaded die?
 - Show that for the loaded die, the probability of obtaining a total of six from four throws is two and half times that of obtaining five, and compare the probability of obtaining 23 with that of obtaining 24 from four throws.
- 9 A fair coin is tossed four times. What is the probability of
- At least three heads?
 - Exactly three heads?
 - A run of three or more consecutive heads?
 - A run of exactly three consecutive heads?
- 10 Find the probability that in 24 throws of two dice, double six fails to appear.
- 11 Two dice are rolled and their scores are denoted by S_1 and S_2 . What is the probability that the quadratic $x^2 + xS_1 + S_2 = 0$ has real roots?
- 12 (a) If $\mathbf{P}(A)$ is the probability that an event A occurs, prove that

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

where A_1, A_2, \dots, A_n are events.

- A tea set consists of six cups and saucers with two cups and saucers in each of three different colours. The cups are placed randomly on the saucers. What is the probability that no cup is on a saucer of the same colour?
- 13 An urn contains three tickets numbered 1, 2, and 3, and they are drawn successively without replacement. What is the probability that there will be at least one value of r ($r = 1, 2, 3$) such that on the r th drawing a ticket numbered r will be drawn?

- 14** Four red balls and two blue balls are placed at random into two urns so that each urn contains three balls. What is the probability of getting a blue ball if
- You select a ball at random from the first urn?
 - You select an urn at random and then select a ball from it at random?
 - You discard two balls from the second urn and select the last ball?
- 15** Four fair dice are rolled and the four numbers shown are multiplied together. What is the probability that this product
- Is divisible by 5?
 - Has last digit 5?
- 16** Suppose that n fair dice are rolled, and let M_n be the product of the numbers shown.
- Show that the probability that the last digit of M_n is 5 is a nonincreasing function of n .
 - Show that the probability that M_n is divisible by 5 is a non-decreasing function of n .
 - Find the limits of the probabilities in (a) and (b) and interpret this.
- 17** The consecutive integers $1, 2, \dots, n$ are inscribed on n balls in an urn. Let D_r be the event that the number on a ball drawn at random is divisible by r .
- What are $\mathbf{P}(D_3)$, $\mathbf{P}(D_4)$, $\mathbf{P}(D_3 \cup D_4)$, and $\mathbf{P}(D_3 \cap D_4)$?
 - Find the limits of these probabilities as $n \rightarrow \infty$.
 - What would your answers be if the n consecutive numbers began at a number $a \neq 1$?
- 18** Show that if A and B are events, then

$$\begin{aligned} \mathbf{P}(A \cap B) - \mathbf{P}(A)\mathbf{P}(B) &= \mathbf{P}(A)\mathbf{P}(B^c) - \mathbf{P}(A \cap B^c) \\ &= \mathbf{P}(A^c)\mathbf{P}(B) - \mathbf{P}(A^c \cap B) \\ &= \mathbf{P}((A \cup B)^c) - \mathbf{P}(A^c)\mathbf{P}(B^c) \end{aligned}$$

- 19** Show that
- $\min\{1, \mathbf{P}(A) + \mathbf{P}(B)\} \geq \mathbf{P}(A \cup B) \geq \max\{\mathbf{P}(A), \mathbf{P}(B)\}$.
 - $\min\{\mathbf{P}(A), \mathbf{P}(B)\} \geq \mathbf{P}(A \cap B) \geq \max\{0, \mathbf{P}(A) + \mathbf{P}(B) - 1\}$.
 - $\mathbf{P}\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{i=1}^n \mathbf{P}(A_i) - (n-1)$.
- 20** The function $d(x, y)$ is defined on the event space by $d(A, B) = \mathbf{P}(A \Delta B)$.
- Show that for any events A, B , and C ,

$$d(A, B) + d(B, C) - d(A, C) = 2(\mathbf{P}(A \cap B^c \cap C) + \mathbf{P}(A^c \cap B \cap C^c)).$$

- When is $d(A, B)$ zero?
- Let A_1, A_2, \dots be a monotone sequence of events such that $A_i \subseteq A_j$ for $i \leq j$. Show that for $i \leq j \leq k$,

$$d(A_i, A_k) = d(A_i, A_j) + d(A_j, A_k).$$

- 21** An urn contains $x \geq 2$ xanthic balls and $y \geq 1$ yellow balls. Two balls are drawn at random without replacement; let p be the probability that both are xanthic.
- If $p = \frac{1}{2}$, find the smallest possible value of x in the two cases when y is odd or even.
 - If $p = \frac{1}{8}$, find the smallest possible value of x .
 - If $p = r^{-2}$, where r is an integer, show that $r \geq 6$, and find values of x and y that yield $p = \frac{1}{36}$.
- 22** When are the following true?
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
 - $A \cap (B \cap C) = (A \cap B) \cap C$
 - $A \cup (B \cup C) = A \setminus (B \setminus C)$
 - $(A \setminus B) \setminus C = A \setminus (B \setminus C)$

- (e) $A \Delta (B \Delta C) = (A \Delta B) \Delta C$
 (f) $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$
 (g) $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$.

- 23 Birthdays** If m students born in 1985 are attending a lecture, show that the probability that at least two of them share a birthday is

$$p = 1 - \frac{(365)!}{(365 - m)!(365)^m}$$

Show that if $m \geq 23$, then $p > \frac{1}{2}$. What difference would it make if they were born in 1988?

- 24** Let $(A_n; n > 1)$ be a collection of events. Show that the event that infinitely many of the A_n occur is given by $\bigcap_{n \geq 1} \bigcup_{m=n}^{\infty} A_m$.

- 25 Boole's Inequality** Show that

$$\mathbf{P}\left(\bigcup_1^n A_i\right) \leq \sum_{i=1}^n \mathbf{P}(A_i).$$

2

Conditional Probability and Independence

Now and then there is a person born who is so unlucky that he runs into accidents which started to happen to somebody else.

Don Marquis

2.1 Conditional Probability

Suppose you have a well-shuffled conventional pack of cards. Obviously (by symmetry), the probability $\mathbf{P}(T)$ of the event T that the top card is an ace is

$$\mathbf{P}(T) = \frac{4}{52} = \frac{1}{13}.$$

However, suppose you notice that the bottom card is the ace of spades S_A . What now is the probability that the top card is an ace? There are 51 possibilities and three of them are aces, so by symmetry again the required probability is $\frac{3}{51}$. To distinguish this from the original probability, we denote it by $\mathbf{P}(T|S_A)$ and call it the conditional probability of T given that the bottom card is the ace of spades.

Similarly, had you observed that the bottom card was the king of spades S_K , you would conclude that the probability that the top card is an ace is

$$\mathbf{P}(T|S_K) = \frac{4}{51}.$$

Here is a less trivial example.

Example: Poker [Note: In this example the symbol $\binom{n}{r}$ denotes the number of ways of choosing r cards from n cards. If you are unfamiliar with this notation, omit this example at a first reading.]

Suppose you are playing poker. As the hand is dealt, you calculate the chance of being dealt a royal flush R , assuming that all hands of five cards are equally likely. (A *royal flush* comprises 10, J, Q, K, A in a single suit.) Just as you get the answer

$$\mathbf{P}(R) = 4 \binom{52}{5}^{-1} = \frac{1}{649740},$$

the dealer deals your last card face up. It is the ace of spades, S_A . If you accept the card, what now is your chance of picking up a royal flush?

Intuitively, it seems unlikely still to be $\mathbf{P}(R)$ above, as the conditions for getting one have changed. Now you need your first four cards to be the ten to king of spades precisely. (Also, had your last card been the two of spades, S_2 , your chance of a royal flush would definitely be zero.) As above, to distinguish this new probability, we call it the conditional probability of R given S_A and denote it by $\mathbf{P}(R|S_A)$.

Is it larger or smaller than $\mathbf{P}(R)$? At least you do have an ace, which is a start, so it might be greater. But you cannot now get a flush in any suit but spades, so it might be smaller. To resolve the uncertainty, you assume that any set of four cards from the remaining 51 cards is equally likely to complete your hand and calculate that

$$\mathbf{P}(R|S_A) = \binom{51}{4}^{-1} = \frac{13}{5}\mathbf{P}(R).$$

Your chances of a royal flush have more than doubled. ●

Let us investigate these ideas in a more general setting. As usual we are given a sample space, an event space \mathcal{F} , and a probability function $\mathbf{P}(\cdot)$. We suppose that some event $B \in \mathcal{F}$ definitely occurs, and denote the conditional probability of any event A , given B , by $\mathbf{P}(A|B)$. As we did for $\mathbf{P}(\cdot)$, we observe that $\mathbf{P}(\cdot|B)$ is a function defined on \mathcal{F} , which takes values in $[0, 1]$. But what function is it?

Clearly, $\mathbf{P}(A)$ and $\mathbf{P}(A|B)$ are not equal in general, because even when $\mathbf{P}(B^c) \neq 0$ we always have

$$\mathbf{P}(B^c|B) = 0.$$

Second, we note that given the occurrence of B , the event A can occur if and only if $A \cap B$ occurs. This makes it natural to require that

$$\mathbf{P}(A|B) \propto \mathbf{P}(A \cap B).$$

Finally, and trivially,

$$\mathbf{P}(B|B) = 1.$$

After a moment's thought about these three observations, it appears that an attractive candidate to play the role of $\mathbf{P}(A|B)$ is $\mathbf{P}(A \cap B)/\mathbf{P}(B)$. We make these intuitive reflections formal as follows.

Definition *Let A and B be events with $\mathbf{P}(B) > 0$. Given that B occurs, the conditional probability that A occurs is denoted by $\mathbf{P}(A|B)$ and defined by*

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

(1)



When $\mathbf{P}(B) = 0$, the conditional probability $\mathbf{P}(A|B)$ is not defined by (1). However, to avoid an endless stream of tiresome reservations about special cases, it is convenient to adopt the convention that, even when $\mathbf{P}(B) = 0$, we may still write $\mathbf{P}(A \cap B) = \mathbf{P}(A|B)\mathbf{P}(B)$, both sides having the value zero. Thus, whether $\mathbf{P}(B) > 0$ or not, it is true

that

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

Likewise, $\mathbf{P}(A \cap B^c) = \mathbf{P}(A|B^c)\mathbf{P}(B^c)$ and hence, for any events A and B , we have proved the following *partition rule*:

(2)	<div style="border: 1px solid black; background-color: #f0f0f0; padding: 10px;"> <p>Theorem $\mathbf{P}(A) = \mathbf{P}(A \cap B) + \mathbf{P}(A \cap B^c)$ $\qquad\qquad = \mathbf{P}(A B)\mathbf{P}(B) + \mathbf{P}(A B^c)\mathbf{P}(B^c).$ ■</p> </div>
-----	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

The reader will come to realize the crucial importance of (1) and (2) as he or she discovers more about probability. We begin with a trivial example.

Example: Poker Revisited Let us check that Definition 1 is consistent with our informal discussion earlier in this section. By (1)

$$\mathbf{P}(R|S_A) = \mathbf{P}(R \cap S_A)/\mathbf{P}(S_A) = \frac{1}{\binom{52}{5}} \bigg/ \frac{\binom{51}{4}}{\binom{52}{5}} = \binom{51}{4}^{-1}. \quad \bullet$$

Here is a more complicated example.

Example: Lemons An industrial conglomerate manufactures a certain type of car in three towns called Farad, Gilbert, and Henry. Of 1000 made in Farad, 20% are defective; of 2000 made in Gilbert, 10% are defective, and of 3000 made in Henry, 5% are defective. You buy a car from a distant dealer. Let D be the event that it is defective, F the event that it was made in Farad and so on. Find: (a) $\mathbf{P}(F|H^c)$; (b) $\mathbf{P}(D|H^c)$; (c) $\mathbf{P}(D)$; (d) $\mathbf{P}(F|D)$. Assume that you are equally likely to have bought any one of the 6000 cars produced.

Solution

$$\begin{aligned} \text{(a)} \quad \mathbf{P}(F|H^c) &= \frac{\mathbf{P}(F \cap H^c)}{\mathbf{P}(H^c)} \quad \text{by (1),} \\ &= \frac{\mathbf{P}(F)}{\mathbf{P}(H^c)} \quad \text{because } F \subseteq H^c, \\ &= \frac{1000}{6000} \bigg/ \frac{3000}{6000} = \frac{1}{3}. \end{aligned}$$

$$\begin{aligned} \text{(b)} \quad \mathbf{P}(D|H^c) &= \frac{\mathbf{P}(D \cap H^c)}{\mathbf{P}(H^c)} \quad \text{by (1)} \\ &= \frac{\mathbf{P}(D \cap (F \cup G))}{\mathbf{P}(H^c)} \quad \text{because } H^c = F \cup G, \\ &= \frac{\mathbf{P}(D \cap F) + \mathbf{P}(D \cap G)}{\mathbf{P}(H^c)} \quad \text{because } F \cap G = \phi \end{aligned}$$

$$\begin{aligned}
&= \frac{\mathbf{P}(D|F)\mathbf{P}(F) + \mathbf{P}(D|G)\mathbf{P}(G)}{\mathbf{P}(H^c)} \quad \text{by (1)} \\
&= \frac{\frac{1}{5} \cdot \frac{1}{6} + \frac{1}{10} \cdot \frac{1}{3}}{\frac{1}{2}} \quad \text{on using the data in the question,} \\
&= \frac{2}{15}.
\end{aligned}$$

$$\begin{aligned}
\text{(c)} \quad \mathbf{P}(D) &= \mathbf{P}(D|H)\mathbf{P}(H) + \mathbf{P}(D|H^c)\mathbf{P}(H^c) \quad \text{by (2)} \\
&= \frac{1}{20} \cdot \frac{1}{2} + \frac{2}{15} \cdot \frac{1}{2} \quad \text{on using the data and (b)} \\
&= \frac{11}{120}.
\end{aligned}$$

$$\begin{aligned}
\text{(d)} \quad \mathbf{P}(F|D) &= \frac{\mathbf{P}(F \cap D)}{\mathbf{P}(D)} \quad \text{by (1)} \\
&= \frac{\mathbf{P}(D|F)\mathbf{P}(F)}{\mathbf{P}(D)} \quad \text{by (1)} \\
&= \frac{\frac{1}{5} \cdot \frac{1}{6}}{\frac{11}{120}} \quad \text{on using the data and (c)} \\
&= \frac{4}{11}.
\end{aligned}$$

We often have occasion to use the following elementary generalization of Theorem 2.

(3) **Theorem** *We have*

$$\mathbf{P}(A) = \sum_i \mathbf{P}(A|B_i)\mathbf{P}(B_i)$$

whenever $A \subseteq \bigcup_i B_i$ *and* $B_i \cap B_j = \phi$ *for* $i \neq j$; *the extended partition rule.*

Proof This is immediate from (1.4.3) and (1). ■

For example, with the notation of (3), we may write

$$\mathbf{P}(B_j|A) = \mathbf{P}(B_j \cap A)/\mathbf{P}(A) = \frac{\mathbf{P}(A|B_j)\mathbf{P}(B_j)}{\mathbf{P}(A)},$$

and expanding the denominator using (3), we have proved the following celebrated result; also known as *Bayes's Rule*:

Bayes's Theorem If $A \subseteq \bigcup_1^n B_i$, and $B_i \cap B_j = \phi$ for $i \neq j$, then

$$(4) \quad \mathbf{P}(B_j|A) = \frac{\mathbf{P}(A|B_j)\mathbf{P}(B_j)}{\sum_1^n \mathbf{P}(A|B_i)\mathbf{P}(B_i)}; \quad \mathbf{P}(A) > 0. \quad \blacksquare$$

The following is a typical example of how (4) is applied in practice.

Example: False Positives You have a blood test for some rare disease that occurs by chance in 1 in every 100 000 people. The test is fairly reliable; if you have the disease, it will correctly say so with probability 0.95; if you do not have the disease, the test will wrongly say you do with probability 0.005. If the test says you do have the disease, what is the probability that this is a correct diagnosis?

Solution Let D be the event that you have the disease and T the event that the test says you do. Then, we require $\mathbf{P}(D|T)$, which is given by

$$\begin{aligned} \mathbf{P}(D|T) &= \frac{\mathbf{P}(T|D)\mathbf{P}(D)}{\mathbf{P}(T|D)\mathbf{P}(D) + \mathbf{P}(T|D^c)\mathbf{P}(D^c)} \quad \text{by (4)} \\ &= \frac{(0.95)(0.00001)}{(0.95)(0.00001) + (0.99999)(0.005)} \simeq 0.002. \end{aligned}$$

Despite appearing to be a pretty good test, for a disease as rare as this the test is almost useless. ●

It is important to note that conditional probability is a probability function in the sense defined in Section 1.4. Thus, $\mathbf{P}(\Omega|B) = 1$ and, if $A_i \cap A_j = \phi$ for $i \neq j$, we have

$$(5) \quad \mathbf{P}\left(\bigcup_i A_i|B\right) = \sum_i \mathbf{P}(A_i|B).$$

From these, we may deduce various useful identities (as we did in Section 1.4); for example:

$$(6) \quad \mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|B \cap C)\mathbf{P}(B|C)\mathbf{P}(C),$$

$$(7) \quad \mathbf{P}\left(\bigcap_1^n A_i\right) = \mathbf{P}\left(A_1|\bigcap_2^n A_i\right)\mathbf{P}\left(A_2|\bigcap_3^n A_i\right)\dots\mathbf{P}(A_n)$$

$$(8) \quad \mathbf{P}(A|B) = 1 - \mathbf{P}(A^c|B),$$

$$(9) \quad \mathbf{P}(A \cup B|C) = \mathbf{P}(A|C) + \mathbf{P}(B|C) - \mathbf{P}(A \cap B|C),$$

and so on.

(10) **Example** Let us prove (5), (6), (7), (8), and (9). First,

$$\begin{aligned} \mathbf{P}\left(\bigcup_i A_i|B\right) &= \mathbf{P}\left(\left(\bigcup_i A_i\right) \cap B\right) / \mathbf{P}(B) \quad \text{by (1).} \\ &= \mathbf{P}\left(\bigcup_i (A_i \cap B)\right) / \mathbf{P}(B) \\ &= \sum_i \mathbf{P}(A_i \cap B) / \mathbf{P}(B) \quad \text{by (1.4.3), because the } A_i \text{ are disjoint,} \\ &= \sum_i \mathbf{P}(A_i|B) \quad \text{by (1) again,} \end{aligned}$$

and we have proved (5). Second, by repeated use of (1),

$$\mathbf{P}(A|B \cap C)\mathbf{P}(B|C)\mathbf{P}(C) = \frac{\mathbf{P}(A \cap B \cap C)}{\mathbf{P}(B \cap C)} \cdot \frac{\mathbf{P}(B \cap C)}{\mathbf{P}(C)} \cdot \mathbf{P}(C) = \mathbf{P}(A \cap B \cap C),$$

if the denominator is not zero. If the denominator is zero, then (6) still holds by convention, both sides taking the value zero.

The relation (7) follows by induction using (6); and (8) and (9) are trivial consequences of (5). ●

(11) **Example: Repellent and Attractive Events** The event A is said to be *attracted* to B if $\mathbf{P}(A|B) > \mathbf{P}(A)$.

If $\mathbf{P}(A|B) < \mathbf{P}(A)$, then A is *repelled* by B and A is *indifferent* to B if

$$(12) \quad \mathbf{P}(A|B) = \mathbf{P}(A).$$

- (a) Show that if B attracts A , then A attracts B , and B^c repels A .
 (b) A flimsy slip of paper is in one of n bulging box files. The event that it is in the j th box file is B_j , where $\mathbf{P}(B_j) = b_j > 0$. The event that a cursory search of the j th box file fails to discover the slip is F_j , where $\mathbf{P}(F_j|B_j) = \phi_j < 1$. Show that B_j and F_j are mutually repellent, but F_j attracts B_i , for $i \neq j$.

Solution (a) Because B attracts A , by (1), $\mathbf{P}(A \cap B) > \mathbf{P}(A)\mathbf{P}(B)$, whence, on dividing by $\mathbf{P}(A)$, we have $\mathbf{P}(B|A) > \mathbf{P}(B)$. Furthermore, by Theorem 2,

$$\begin{aligned} \mathbf{P}(A|B^c)\mathbf{P}(B^c) &= \mathbf{P}(A) - \mathbf{P}(A|B)\mathbf{P}(B) < \mathbf{P}(A)(1 - \mathbf{P}(B)), \quad \text{because } B \text{ attracts } A, \\ &= \mathbf{P}(A)\mathbf{P}(B^c). \end{aligned}$$

So B^c repels A (on dividing through by $\mathbf{P}(B^c) \neq 0$).

(b) By Bayes' theorem (4),

$$\mathbf{P}(B_j|F_j) = \frac{\mathbf{P}(F_j|B_j)\mathbf{P}(B_j)}{\sum_{i=1}^n \mathbf{P}(F_j|B_i)\mathbf{P}(B_i)} = \frac{\phi_j b_j}{1 - b_j + \phi_j b_j}$$

because, obviously, for $i \neq j$, $\mathbf{P}(F_j|B_i) = 1$. Hence,

$$\mathbf{P}(B_j) - \mathbf{P}(B_j|F_j) = \frac{b_j(1-b_j)(1-\phi_j)}{1-b_j+\phi_j b_j} > 0.$$

Therefore, B_j is repelled by F_j . Also, for $i \neq j$,

$$\mathbf{P}(B_i|F_j) - \mathbf{P}(B_i) = \frac{b_i}{1-b_j+\phi_j b_j} - b_i = \frac{b_i b_j (1-\phi_j)}{1-b_j+\phi_j b_j} > 0$$

so F_j attracts B_i , for $i \neq j$. ●

Notice that this agrees with our intuition. We believe quite strongly that if we look in a file for a slip and fail to find it, then it is more likely (than before the search) to be elsewhere. (Try to think about the consequences if the opposite were true.) This conclusion of Example 11 was not incorporated in our axioms, but follows from them. It therefore lends a small but valuable boost to their credibility.

Finally, we consider sequences of conditional probabilities. Because conditional probability is a probability function [see (5)], we expect it to be continuous in the sense of Section 1.5. Thus if (as $n \rightarrow \infty$) $A_n \rightarrow A$ and $B_n \rightarrow B$, then by Theorem 1.5.2 we have

$$\lim_{n \rightarrow \infty} \mathbf{P}(A_n|B) = \mathbf{P}(A|B)$$

and

$$(13) \quad \lim_{n \rightarrow \infty} \mathbf{P}(A|B_n) = \mathbf{P}(A|B).$$

2.2 Independence

It may happen that the conditional probability $\mathbf{P}(A|B)$ is the same as the unconditional probability $\mathbf{P}(A)$, so that

$$\mathbf{P}(A) = \mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

This idea leads to the following:

(1) **Definition** (a) Events A and B are independent when

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

(b) A collection of events $(A_i; i \geq 1)$ is independent when

$$\mathbf{P}\left(\bigcap_{i \in F} A_i\right) = \prod_{i \in F} \mathbf{P}(A_i)$$

for any finite set F of indices.

(c) Events A and B are conditionally independent, given C , when

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C).$$

This does not imply independence unless $C = \Omega$.

(d) A collection of events $(A_i; i \geq 1)$ is pairwise independent if

$$\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j) \quad \text{for } i \neq j.$$

This does not imply independence in general. ▲

It is easy to see that independence is equivalent to the idea of indifference defined in (2.1.12), but the term “indifference” is not in general use. It is usually, but not always, clear when two events are independent, as the next two examples illustrate.

(2) **Example: Sport** Prior to a game of football, you toss a coin for the kick-off. Let C be the event that you win the toss, and let M be the event that you win the match.

(a) Show that the outcome of the match is independent of whether you win the toss if and only if, for some p and p' , with $0 < p, p' < 1$,

$$\begin{aligned}\mathbf{P}(C \cap M) &= pp', \\ \mathbf{P}(C \cap M^c) &= p(1 - p'), \\ \mathbf{P}(C^c \cap M) &= (1 - p)p',\end{aligned}$$

and

$$\mathbf{P}(C^c \cap M^c) = (1 - p)(1 - p').$$

(b) Let B be the event that you win both or lose both, so $B = \{(C \cap M) \cup (C^c \cap M^c)\}$. Suppose that C and M are indeed independent. Show that C and B are independent if and only if $p' = \frac{1}{2}$.

Solution (a) If C and M are independent, and $\mathbf{P}(C) = p$ and $\mathbf{P}(M) = p'$, then by definition $\mathbf{P}(C \cap M) = pp'$ and so on.

Conversely, for the given probabilities

$$\mathbf{P}(C) = \mathbf{P}(C \cap M) + \mathbf{P}(C \cap M^c) = pp' + p(1 - p') = p$$

and similarly we have $\mathbf{P}(M) = p'$. Hence,

$$\mathbf{P}(C)\mathbf{P}(M) = pp' = \mathbf{P}(C \cap M).$$

This, together with three similar identities (exercises for you), demonstrates the independence.

(b) Trivially, $\mathbf{P}(C \cap B) = \mathbf{P}(C \cap M)$. Hence, C and B are independent if

$$pp' = \mathbf{P}(C \cap M) = \mathbf{P}(C)\mathbf{P}(B) = p(pp' + (1 - p)(1 - p')).$$

That is, if $(1 - p)(1 - 2p') = 0$. Because $p \neq 1$, it follows that $p' = \frac{1}{2}$. The converse is trivial. ●

(3) **Example: Flowers** A plant gets two independent genes for flower colour, one from each parent plant. If the genes are identical, then the flowers are uniformly of that colour;

if they are different, then the flowers are striped in those two colours. The genes for the colours pink, crimson, and red occur in the population in the proportions $p:q:r$, where $p + q + r = 1$. A given plant's parents are selected at random; let A be the event that its flowers are at least partly pink, and let B be the event that its flowers are striped.

- (a) Find $\mathbf{P}(A)$ and $\mathbf{P}(B)$.
 (b) Show that A and B are independent if $p = \frac{2}{3}$ and $r = q = \frac{1}{6}$.
 (c) Are these the only values of p , q , and r such that A and B are independent?

Solution (a) With an obvious notation (P for pink, C for crimson, and R for red), we have

$$\begin{aligned}\mathbf{P}(PP) &= \mathbf{P}(P)\mathbf{P}(P), \quad \text{by parents independence,} \\ &= p^2,\end{aligned}$$

because P occurs with probability p . Likewise,

$$\mathbf{P}(PR) = \mathbf{P}(R)\mathbf{P}(P) = rp = \mathbf{P}(RP).$$

Hence,

$$\begin{aligned}\mathbf{P}(A) &= \mathbf{P}(PP \cup PR \cup P \cup PC \cup CP) \\ &= p^2 + 2pr + 2pq \quad \text{by (1.4.3),} \\ &= 1 - (1 - p)^2,\end{aligned}$$

because $p + q + r = 1$. (Can you see how to get this last expression directly?) Similarly,

$$\mathbf{P}(B) = \mathbf{P}(PC \cup PR \cup RC) = 2(pq + qr + rp).$$

- (b) The events A and B are independent, if and only if,

$$\mathbf{P}(A)\mathbf{P}(B) = \mathbf{P}(A \cap B) = \mathbf{P}(PC \cup PR) = 2(pq + pr).$$

From part (a), this is equivalent to

$$(4) \quad (1 - (1 - p)^2)(pq + qr + pr) = p(q + r),$$

and this is satisfied by the given values of p , q , and r .

(c) No. Rearranging (4), we see that A and B are independent for any values of q and r lying on the curve $rq = 2rq(q + r) + r^3 + q^3$, in the $r - q$ plane. You may care to amuse yourself by showing that this is a loop from the origin. Outside the loop, A and B are attractive; inside the loop, A and B are repellent. ●

- (5) **Example 1.13 Revisited: Craps** Let us reconsider this game using conditional probability and independence. Recall that A_k is the event that you win by rolling a pair with sum k . Let S_k be the event that any given roll yields sum k . Now, for example, A_4 occurs only if S_4 occurs at the first roll and S_4 occurs before S_7 in later rolls. However, all the rolls after the first until the first occurrence of S_4 or S_7 are irrelevant, and rolls are independent.

Hence,

$$\mathbf{P}(A_4) = \mathbf{P}(S_4)\mathbf{P}(S_4|S_4 \cup S_7) = \frac{(\mathbf{P}(S_4))^2}{\mathbf{P}(S_4 \cup S_7)} = \frac{\left(\frac{3}{36}\right)^2}{\frac{3}{36} + \frac{6}{36}} = \frac{1}{36}.$$

Now, performing a similar calculation for A_5 , A_6 , A_8 , A_9 , and A_{10} yields the solution to Example 1.13. ●

(6) Example Suppose A and B are independent, and B and C are independent.

- (a) Are A and C independent in general?
- (b) Is B independent of $A \cup C$?
- (c) Is B independent of $A \cap C$?

Solution (a) No. Take $A \equiv C$. [Then do Problem 4(a).]

(b) No. Consider Example 2(b), with $p = p' = \frac{1}{2}$, and let $A = M$. Then, B is independent of M and C , but

$$\mathbf{P}(B \cap (C \cup M)) = \mathbf{P}(C \cap M) = \frac{1}{4}$$

and

$$\mathbf{P}(B)\mathbf{P}(C \cup M) = (pp' + (1 - p)(1 - p'))(1 - (1 - p')(1 - p)) = \frac{3}{8} \neq \frac{1}{4}.$$

(c) No. With the same notation as in (b), following Example 2(b) again, we have

$$\mathbf{P}(B \cap C \cap M) = \mathbf{P}(C \cap M) = \frac{1}{4},$$

and

$$\mathbf{P}(B)\mathbf{P}(C \cap M) = (pp' + (1 - p')(1 - p))pp' = \frac{1}{8} \neq \frac{1}{4}. \quad \bullet$$

2.3 Recurrence and Difference Equations

Many problems in probability have a structure in which the repetition of some procedure is essential. At a trivial level, one may repeatedly roll dice, catch fish, have children, and so on; more important problems involve the same idea. In Chapter 1, it is necessary to suppose that all the probabilities on the sample space were given or that all outcomes were equally likely. Conditional probability provides a more natural way of defining such problems; conditional on the procedure having reached some stage, it supplies the probabilities of the consequent events.

By emphasizing this recurrent aspect of some experiment, conditional probability enables us to tackle problems by deriving recurrence relations. These often turn out to be simple difference equations. Many of the worked examples illustrate these ideas; the following results are useful.

(1) **Theorem** Let $(a_k; k \geq 0)$ be a sequence of real numbers.

(i) If the sequence $(u_k; k \geq 0)$ satisfies

$$u_{k+1} - a_k u_k = 0,$$

then

$$u_k = u_0 \prod_0^{k-1} a_j.$$

(ii) If $(u_k; k \geq 0)$ satisfies

$$u_{k+1} - u_k = c\alpha^k,$$

where α and c are constants, then

$$u_k - u_0 = c \frac{\alpha^k - 1}{\alpha - 1}.$$

(iii) If $(u_k; k \geq 0)$ satisfies

$$u_{k+1} - u_k = c,$$

for some constant c , then

$$u_k - u_0 = kc.$$

(iv) If for some constants a, α, c_1 , and c_2 , we have

$$u_{k+1} - au_k = c_1 + c_2\alpha^k, a \neq \alpha,$$

then

$$u_k = u_0 a^k + \frac{c_1(1 - a^k)}{1 - a} + \frac{c_2(\alpha^k - a^k)}{\alpha - a}.$$

Proof These results are verified simply by substitution. ■

(2) **Example: Fly** A room has four walls, a floor, and a ceiling. A fly moves between these surfaces. If it leaves the floor or ceiling, then it is equally likely to alight on any one of the four walls or the surface it has just left. If it leaves a wall, then it is equally likely to alight on any of the other three walls, the floor, or the ceiling. Initially, it is on the ceiling.

Let F_k denote the event that it is on the floor after k moves. What is $f_k = \mathbf{P}(F_k)$?

Solution Let C_k denote the event that it is on the ceiling after k moves, and N_k, E_k, W_k, S_k denote the corresponding event for the four walls. Set $c_k = \mathbf{P}(C_k)$, and so on. Then by Theorem 2.1.3,

$$\begin{aligned} (3) \quad \mathbf{P}(F_k) &= \mathbf{P}(F_k|F_{k-1})\mathbf{P}(F_{k-1}) + \mathbf{P}(F_k|C_{k-1})\mathbf{P}(C_{k-1}) + \mathbf{P}(F_k|N_{k-1})\mathbf{P}(N_{k-1}) \\ &\quad + \mathbf{P}(F_k|E_{k-1})\mathbf{P}(E_{k-1}) + \mathbf{P}(F_k|W_{k-1})\mathbf{P}(W_{k-1}) + \mathbf{P}(F_k|S_{k-1})\mathbf{P}(S_{k-1}) \\ &= \frac{1}{5}f_{k-1} + 0 + \frac{4}{5}w_{k-1}, \end{aligned}$$

where we have used the fact that, by symmetry, $n_k = e_k = w_k = s_k$. Likewise,

$$w_k = \frac{3}{5}w_{k-1} + \frac{1}{5}f_{k-1} + \frac{1}{5}c_{k-1} = \frac{3}{5}w_{k-1} + \frac{1}{5}(1 - 4w_{k-1}),$$

on using the fact that

$$(4) \quad f_k + c_k + 4w_k = 1,$$

which follows from the observation that the fly has to be somewhere in the room. Hence,

$$(5) \quad w_k = -\frac{1}{5}w_{k-1} + \frac{1}{5}.$$

Because the fly starts on the ceiling, $w_0 = 0$, and so by Theorem 1(iv)

$$(6) \quad w_k = \frac{1}{6} \left(1 - \left(-\frac{1}{5} \right)^k \right).$$

Substituting into (3) gives

$$f_k = \frac{1}{5}f_{k-1} + \frac{2}{15} \left(1 - \left(-\frac{1}{5} \right)^{k-1} \right).$$

Hence, for some constant A ,

$$f_k = A \left(\frac{1}{5} \right)^k + \frac{1}{3} \left(-\frac{1}{5} \right)^k + \frac{1}{6}.$$

The arbitrary constant A is determined by recalling that the fly starts on the ceiling. Thus, $f_0 = 0$, and

$$(7) \quad f_k = -\frac{1}{2} \left(\frac{1}{5} \right)^k + \frac{1}{3} \left(-\frac{1}{5} \right)^k + \frac{1}{6}.$$

Alternatively, we may substitute (5) into (3) to get directly:

$$(8) \quad f_k = \frac{1}{25}f_{k-2} + \frac{4}{25}.$$

It is a simple matter to check that, for any constants A and B ,

$$f_k = A \left(\frac{1}{5} \right)^k + B \left(-\frac{1}{5} \right)^k + \frac{1}{6}$$

satisfies (8). Because $f_0 = f_1 = 0$, the solution (7) is recovered immediately. Notice that as $k \rightarrow \infty$, (6) and (7) yield $f_k \rightarrow \frac{1}{6}$ and $w_k \rightarrow \frac{1}{6}$. It follows from (4) that $c_k \rightarrow \frac{1}{6}$. In the long run, the fly is equally likely to be on any surface. ●

2.4 Remarks

Independence and conditioning greatly add to your armoury of weapons for attacking problems. If an event requires the occurrence of a number of independent events, then calculations are simplified by using Definition 2.2.1.

Where independence fails, it may be possible to find a family of disjoint events B_i whose union includes the event A of interest, and for which $\mathbf{P}(A|B_i)$ is easily calculated. The required probability is then calculated using Theorem 2.1.3.

Such events B_i can also be found to yield $\mathbf{P}(A)$ as the solution of some recurrence relation, as exemplified in Section 2.3.

Note that you are warned to avoid the painful student error that asserts that A and B are independent when $A \cap B = \phi$. This is wrong, except possibly when one of $\mathbf{P}(A)$ or $\mathbf{P}(B)$ is zero, which is not a case of any great interest in general.

Finally, we take this opportunity to stress that although conditional probability is extremely useful and powerful, it also greatly increases the possibilities for making egregious errors. The celebrated Monty Hall problem is a recent classic example, which we discuss in Example 2.13. But cautionary examples had been in existence for many years before Marilyn vos Savant made that one famous. Here are two classics from the nineteenth century.

Galton's Paradox (1894) Suppose you flip three fair coins. At least two are alike and it is an evens chance whether the third is a head or a tail, so the chance that all three are the same is $\frac{1}{2}$.

Solution In fact,

$$\mathbf{P}(\text{all same}) = \mathbf{P}(TTT) + \mathbf{P}(HHH) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.$$

What is wrong?

Resolution As is often the case, this paradox arises from fudging the sample space. This "third" coin is not identified initially in Ω , it is determined by the others. The chance whether the "third" is a head or a tail is a conditional probability, not an unconditional probability. Easy calculations show that

$$\left. \begin{array}{l} \mathbf{P}(\text{3rd is } H|HH) = \frac{1}{4} \\ \mathbf{P}(\text{3rd is } T|HH) = \frac{1}{4} \\ \mathbf{P}(\text{3rd is } T|TT) = \frac{1}{4} \\ \mathbf{P}(\text{3rd is } H|TT) = \frac{1}{4} \end{array} \right\} \begin{array}{l} HH \text{ denotes the event that there} \\ \text{are at least two heads.} \\ \\ TT \text{ denotes the event that there} \\ \text{are at least two tails.} \end{array}$$

In *no circumstances*, therefore, is it true that it is an evens chance whether the "third" is a head or a tail; the argument collapses. ●

Bertrand's Box Paradox (1889) There are three boxes. One contains two black counters, one contains two white counters, and one contains a black and a white counter. Pick a box at random and remove a counter without looking at it; it is equally likely to be black or white. The other counter is equally likely to be black or white. Therefore, the chance that your box contains identical counters is $\frac{1}{2}$. However, this is clearly false: the correct answer is $\frac{2}{3}$.

Resolution This is similar to Galton's paradox. Having picked a box and counter, the probability that the other counter is the same is a conditional probability, not an unconditional probability. Thus, easy calculations give (with an obvious notation)

$$(1) \quad \mathbf{P}(\text{both black}|B) = \frac{2}{3} = \mathbf{P}(\text{both white}|W);$$

in neither case is it true that the *other* counter is equally likely to be black or white. ●

2.5 Review and Checklist for Chapter 2

In practice, we may have partial knowledge about the outcome of an experiment, or the conditions of an experiment may change. We therefore used intuitive ideas about probability to define conditional probability and deduce the key result known as Bayes's rule (or Bayes's theorem). It may also be the case that occurrence of an event has no effect on the probability of another. This led us to define the concept of independence.

These ideas are particularly useful when experiments have (or can be reformulated to have) a sequential structure. Use of conditional probability and independence often gives rise to recurrence relations and difference equations in these cases.

Notation: $\mathbf{P}(A|B)$ The conditional probability of A given B .

RULES:

Conditioning Rule: $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$

Addition Rule: $\mathbf{P}(A \cup B|C) = \mathbf{P}(A|C) + \mathbf{P}(B|C)$, when $A \cap C$ and $B \cap C$ are disjoint.

Multiplication Rule: $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|B \cap C)\mathbf{P}(B|C)\mathbf{P}(C)$

Partition Rule: $\mathbf{P}(A) = \sum \mathbf{P}(A|B_i)\mathbf{P}(B_i)$, when $(B_i : i \geq 1)$ are disjoint events and $A \subseteq \cup_i B_i$.

Bayes's Rule (or Theorem):

$$\mathbf{P}(B_i|A) = \mathbf{P}(A|B_i)\mathbf{P}(B_i)/\mathbf{P}(A) = \frac{\mathbf{P}(A|B_i)\mathbf{P}(B_i)}{\sum_j \mathbf{P}(A|B_j)\mathbf{P}(B_j)}$$

Extended addition Rule:

$$\mathbf{P}(\cup_i A_i|C) = \sum_i \mathbf{P}(A_i|C)$$

when $\{A_i\}$ is a partition; which is to say $A_i \cap A_j \neq \phi$, and $C \subseteq \cup_i A_i$.

Extended multiplication Rule:

$$\mathbf{P}\left(\bigcap_{i=1}^n A_i\right) = \mathbf{P}(A_1|A_2 \cap \dots \cap A_n) \dots \mathbf{P}(A_{n-1}|A_n)\mathbf{P}(A_n)$$

Independence Rule: A and B are independent if and only if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$; this is equivalent to $\mathbf{P}(A|B) = \mathbf{P}(A)$ and to $\mathbf{P}(B|A) = \mathbf{P}(B)$.

More generally, $(A_i; 1 \leq i \leq n)$ are independent if $\mathbf{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbf{P}(A_i)$ for any

choice of index set $I \subseteq \{1, \dots, n\}$.

Conditional independence Rule: A and B are conditionally independent given C when

$$\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C).$$

Note that independence does not imply conditional independence, nor is it implied by it.

Pairwise independence Rule: $(A_i; 1 \leq i \leq n)$ are pairwise independent if

$$\mathbf{P}(A_i \cap A_j) = \mathbf{P}(A_i)\mathbf{P}(A_j), \quad i \neq j.$$

Checklist of Terms for Chapter 2

- 2.1 conditional probability
 - partition rules
 - Bayes's theorem (or rule)
- 2.2 independence
 - conditional independence
 - pairwise independence
- 2.3 difference equation
- 2.4 paradoxes

WORKED EXAMPLES AND EXERCISES

2.6 Example: Sudden Death

Two golfers (Atropos and Belladonna) play a series of holes. Atropos wins each hole with probability p , Belladonna wins each hole with probability q , and holes are halved with probability r . Holes are independent, and the game stops on the first occasion when a hole is not halved. What is the probability that Atropos wins?

We give two methods of solution.

Solution I Let A_n be the event that Atropos wins the match at the n th hole, H_k the event that the k th hole is halved, and W_k the event that Atropos wins the k th hole. Then,

$$(1) \quad \mathbf{P}(A_n) = \mathbf{P}(H_1 \cap H_2 \cap \dots \cap H_{n-1} \cap W_n) = \prod_1^{n-1} \mathbf{P}(H_k)\mathbf{P}(W_n) \quad \text{by independence}$$

$$= r^{n-1}p.$$

Hence, the probability that Atropos wins the match is

$$(2) \quad \mathbf{P}\left(\bigcup_1^{\infty} A_n\right) = \sum_1^{\infty} \mathbf{P}(A_n)$$

by (1.4.3) because $A_k \cap A_j = \phi$ for $k \neq j$. Now

$$\sum_1^{\infty} \mathbf{P}(A_n) = \sum_{n=1}^{\infty} r^{n-1}p = \frac{p}{1-r} = \frac{p}{p+q}.$$

Solution II Let V be the event that Atropos wins the match. Then by Theorem 2.1.3,

$$(3) \quad \mathbf{P}(V) = \mathbf{P}(V|W_1)\mathbf{P}(W_1) + \mathbf{P}(V|H_1)\mathbf{P}(H_1) + \mathbf{P}(V|W_1^c \cap H_1^c)\mathbf{P}(W_1^c \cap H_1^c).$$

Now, $\mathbf{P}(V|W_1) = 1$, and

$$\mathbf{P}(V|H_1^c \cap W_1^c) = 0.$$

Also, because holes are independent

$$\mathbf{P}(V|H_1) = \mathbf{P}(V).$$

Hence, substituting into (3),

$$\mathbf{P}(V) = p + \mathbf{P}(V)r,$$

so $\mathbf{P}(V) = p/(1 - r)$, in agreement with Solution I.

Remark The first solution harks back to Chapter 1, by dividing up the sample space into disjoint events and using (1.4.3). The second solution exploits the power of conditional probability by conditioning on the outcome of the first hole. You will use this second idea frequently in tackling problems in probability.

- (4) **Exercise** Show that the probability u_n that Atropos wins at or before the n th hole is $p(1 - r^n)/(1 - r)$.
- (5) **Exercise** Given that Atropos wins at or before the n th hole, show that:
- The probability that the first hole was halved is $r(1 - r^{n-1})/(1 - r^n)$,
 - The probability that the first hole was won is $(1 - r)/(1 - r^n)$.
- (6) **Exercise** Given that Atropos wins, what is the probability that she has won before the third hole?
- (7) **Exercise** What is the probability that Atropos wins, given that exactly n holes have been played when the match is won? Use this to solve the example by a third method.

2.7 Example: Polya's Urn

An urn contains b blue balls and c cyan balls. A ball is drawn at random, its colour is noted, and it is returned to the urn together with d further balls of the same colour. This procedure is repeated indefinitely. What is the probability that:

- The second ball drawn is cyan?
- The first ball drawn is cyan given that the second ball drawn is cyan?

Solution Let C_n denote the event that the n th drawn ball is cyan. Then

$$(a) \quad \mathbf{P}(C_2) = \mathbf{P}(C_2|C_1)\mathbf{P}(C_1) + \mathbf{P}(C_2|C_1^c)\mathbf{P}(C_1^c).$$

Now given C_1 , the urn contains $c + d$ cyan balls on the second drawing, so

$$\mathbf{P}(C_2|C_1) = \frac{c + d}{b + c + d}.$$

Likewise, given C_1^c the urn contains c cyan balls on the second drawing, so

$$\mathbf{P}(C_2|C_1^c) = \frac{c}{b+c+d}.$$

Hence,

$$\mathbf{P}(C_2) = \frac{c+d}{b+c+d} \cdot \frac{c}{b+c} + \frac{c}{b+c+d} \cdot \frac{b}{b+c} = \frac{c}{b+c} = \mathbf{P}(C_1).$$

$$(b) \quad \mathbf{P}(C_1|C_2) = \mathbf{P}(C_1 \cap C_2)\mathbf{P}(C_2) = \mathbf{P}(C_2|C_1)\mathbf{P}(C_1)/\mathbf{P}(C_2) = \frac{c+d}{b+c+d}.$$

using the results of (a).

- (1) **Exercise** Show that $\mathbf{P}(C_n) = \mathbf{P}(C_1)$ for all $n \geq 1$.
- (2) **Exercise** Find the probability that the first drawn ball is cyan given that the n th drawn ball is cyan.
- (3) **Exercise** Find the probability that the first drawn ball is cyan given that the following n drawn balls are all cyan. What is the limit of this probability as $n \rightarrow \infty$?
- (4) **Exercise** Show that for any j, k , $\mathbf{P}(C_k|C_j) = \mathbf{P}(C_j|C_k)$.
- (5) **Exercise** Show that in $m+n$ drawings, the probability that m cyan balls are followed by n blue balls is the same as the probability that n blue balls are followed by m cyan balls. Generalize this result.

2.8 Example: Complacency

In a factory, if the most recent accident occurred exactly k days before today, then the probability that an accident occurs today is p_k ; there is no accident with probability $1 - p_k$. During the n successive days immediately after an accident, what is the probability that

- (a) There are no accidents?
- (b) There is exactly one accident?

Solution (a) Let A_n be the event that the n days following an accident are free of accidents, $n \geq 1$. We are given that

$$(1) \quad \mathbf{P}(A_n|A_{n-1}) = 1 - p_n, \quad n \geq 2$$

and $\mathbf{P}(A_1) = 1 - p_1$. The crucial observation is that

$$(2) \quad \mathbf{P}(A_n) = \mathbf{P}(A_n|A_{n-1})\mathbf{P}(A_{n-1}) = (1 - p_n)\mathbf{P}(A_{n-1}).$$

This is almost completely obvious, but we labour the point by giving two detailed verifications.

I Notice that $A_n \subseteq A_{n-1}$. Hence,

$$\mathbf{P}(A_n) = \mathbf{P}(A_n \cap A_{n-1}) = \mathbf{P}(A_n|A_{n-1})\mathbf{P}(A_{n-1}) \quad \text{by (2.1.1).}$$

II Alternatively, by Theorem 2.1.2,

$$\mathbf{P}(A_n) = \mathbf{P}(A_n|A_{n-1})\mathbf{P}(A_{n-1}) + \mathbf{P}(A_n|A_{n-1}^c)\mathbf{P}(A_{n-1}^c) = \mathbf{P}(A_n|A_{n-1})\mathbf{P}(A_{n-1})$$

because $\mathbf{P}(A_n|A_{n-1}^c) = 0$.

Returning to (1), we iterate this relation to get

$$\mathbf{P}(A_n) = \prod_1^n (1 - p_j) = \pi_n \quad (\text{say}).$$

(b) Let B_k be the event that the first accident in the n day period occurs on the k th day. Then,

$$\begin{aligned} \mathbf{P}(B_k) &= \mathbf{P}(A_k^c \cap A_{k-1}) = \mathbf{P}(A_k^c | A_{k-1}) \mathbf{P}(A_{k-1}) \quad \text{by (2.1.1)} \\ &= p_k \prod_1^{k-1} (1 - p_j) = p_k \pi_{k-1} \end{aligned}$$

Now, given an accident on the k th day, the event that there are no accidents in the succeeding $n - k$ days has the same probability as A_{n-k} . Hence, the probability of exactly one accident is

$$\sum_{k=1}^n \mathbf{P}(B_k) \mathbf{P}(A_{n-k}) = \sum_{k=1}^n \pi_{k-1} p_k \pi_{n-k} = s_n \quad (\text{say}).$$

- (3) **Exercise** Show that if p_n is nondecreasing in n (and $p_n > 0$), then an accident is certain to occur sometime.
- (4) **Exercise** Evaluate π_n and s_n when $p_n = p$.
- (5) **Exercise** What is the probability that in the n days following an accident
- There is at least one accident?
 - There are exactly two accidents?
 - Evaluate these probabilities when $p_n = p$.
- (6) **Exercise** Show that if c_i is a collection of numbers satisfying $0 \leq c_i \leq 1, i \geq 1$, then

$$c_1 + \sum_{i=2}^{\infty} c_i \prod_{j=1}^{i-1} (1 - c_j) + \prod_{i=1}^{\infty} (1 - c_i) = 1.$$

- (7) **Exercise** What condition on $(p_n; n \geq 1)$ would allow a nonzero chance of no second accident?

2.9 Example: Dogfight

Algy, Biggles, and the Commodore are fighting an air battle. In the first round, each fires one shot in alphabetical order, and each may fire at any unhit opponent. Anyone hit drops out of the battle immediately. Survivors play successive rounds with the same rules as the first round until one winner remains.

On any shot aimed at an opponent, Algy hits with probability α , the Commodore hits with probability γ , and Biggles never misses. Show that if shots are independent and

$$\gamma > \alpha > 1 - \frac{\gamma}{1 - \gamma},$$

then Algy should fire his first shot into the ground.

Solution Suppose that Algy were to fire at Biggles and hit him. Algy would then be the Commodore's target, and the battle would continue with shots alternating between these two until a hit is scored. Let C_A be the event that the Commodore wins this two-man battle, and consider the following three events:

$A_1 \equiv$ the Commodore scores a hit with his first shot.

$A_2 \equiv$ the Commodore misses and Algy scores a hit with his first returned shot.

$A_3 \equiv$ the Commodore and Algy both miss their first shots.

If A_3 occurs, then the next round begins under the same conditions; hence,

$$\mathbf{P}(C_A|A_3) = \mathbf{P}(C_A).$$

Also, $\mathbf{P}(C_A|A_1) = 1$, and $\mathbf{P}(C_A|A_2) = 0$. Because $A_i \cap A_j = \phi$ for $i \neq j$ and $\cup_1^3 A_i = \Omega$, we may use Theorem 2.1.3 to give

$$\mathbf{P}(C_A) = \gamma + 0 + (1 - \gamma)(1 - \alpha)\mathbf{P}(C_A),$$

yielding

$$\mathbf{P}(C_A) = \frac{\gamma}{1 - (1 - \gamma)(1 - \alpha)}.$$

However, if Algy misses, then Biggles will certainly fire at the Commodore because $\alpha < \gamma$, and hit him. Then Algy can win only if his second round shot at Biggles hits; otherwise, Biggles surely hits him with his second round shot. Thus, in this case, Algy wins with probability α .

Hence, missing Biggles gives Algy a better chance if

$$1 - \mathbf{P}(C_A) = \frac{\alpha(1 - \gamma)}{1 - (1 - \gamma)(1 - \alpha)} > \alpha$$

that is if $\alpha > 1 - \gamma/(1 - \gamma)$.

- (1) **Exercise** If Algy does fire his first shot at Biggles, what is the probability that he wins the battle?
- (2) **Exercise** Algy is not a competent probabilist, and decides whether to shoot at Biggles by tossing a coin (heads he does, tails he doesn't). Given that the battle ends with the fourth shot, what is the probability that Algy aimed to miss?
- (3) **Exercise** Suppose that Biggles is not infallible; in fact, the probability that any shot of his hits is β . If $\alpha = 0.5$, $\beta = 0.875$, and $\gamma = 0.75$, where should Algy aim his first shot?

2.10 Example: Smears

In a population of women, a proportion p have abnormal cells on the cervix. The Pap test entails taking a sample of cells from the surface of the cervix and examining the sample to detect any abnormality.

- (i) In a case where abnormal cells are present, the sample will fail to include any with probability μ .
- (ii) In a sample including abnormal cells, examination fails to observe them with probability ν .
- (iii) In a sample free of abnormal cells, normal cells are wrongly classified as abnormal with probability π .

All sampling and identification errors are independent.

If a randomly selected woman has such a test:

- (a) What is the probability that the result is wrong?
- (b) If an abnormality is reported, what is the probability that no abnormal cells are present?

Solution (a) Let E denote the event that the result is in error, A denote the event that abnormal cells are present, and S denote the event that the sample fails to include abnormal cells when they are present. Using Theorem 2.1.2, we have

$$\mathbf{P}(E) = \mathbf{P}(E|A)\mathbf{P}(A) + \mathbf{P}(E|A^c)\mathbf{P}(A^c) = \mathbf{P}(E|A)p + \mathbf{P}(E|A^c)(1 - p).$$

By (iii), $\mathbf{P}(E|A^c) = \pi$. Also, by (2.1.5),

$$\mathbf{P}(E|A) = \mathbf{P}(E \cap S|A) + \mathbf{P}(E \cap S^c|A).$$

By (iii) and (i),

$$\mathbf{P}(E \cap S|A) = (1 - \pi)\mu$$

and by (ii) and (i),

$$\mathbf{P}(E \cap S^c|A) = \nu(1 - \pi).$$

Hence,

$$\mathbf{P}(E) = p((1 - \pi)\mu + \nu(1 - \mu)) + (1 - p)\pi.$$

(b) Let D denote the event that an abnormality is reported. By (2.1.1),

$$\mathbf{P}(A^c|D) = \frac{\mathbf{P}(A^c \cap D)}{\mathbf{P}(D)}.$$

Now, by Theorem 2.1.2,

$$\begin{aligned} \mathbf{P}(D) &= \mathbf{P}(D|A)\mathbf{P}(A) + \mathbf{P}(D|A^c)\mathbf{P}(A^c) \\ &= \mathbf{P}(D \cap S|A)\mathbf{P}(A) + \mathbf{P}(D \cap S^c|A)\mathbf{P}(A) + \pi(1 - p) \quad \text{by (2.1.5)} \\ &= \pi\mu p + (1 - \nu)(1 - \mu)p + \pi(1 - p) \end{aligned}$$

and

$$\mathbf{P}(A^c \cap D) = \pi(1 - p).$$

Hence,

$$\mathbf{P}(A^c|D) = \frac{\pi(1 - p)}{\pi\mu p + \pi(1 - p) + (1 - \nu)(1 - \mu)p}.$$

Notice that this is an example of Bayes' Theorem (2.1.4).

- (1) **Exercise** Evaluate $\mathbf{P}(E)$ and $\mathbf{P}(A^c|D)$ when
 - (a) $p = 10\%$ and $\mu = \nu = \pi = 10^{-1}$, and when
 - (b) $p = 50\%$ and $\mu = 10^{-1}$ and $\nu = \pi = 10^{-2}$.
- (2) **Exercise** What is the probability that the result is wrong if no abnormality is reported? Evaluate this in the above two cases, and compare $\mathbf{P}(A^c|D)$ and $\mathbf{P}(A|D^c)$.
- (3) **Exercise** Whatever the result of the test, it is recorded wrongly in the letter to the patient with probability ρ independently of other errors. Let L be the event that the letter is wrong, and let M be the event that the letter reports abnormalities to the patient. Find $\mathbf{P}(L)$, $\mathbf{P}(A^c|M)$, and $\mathbf{P}(A|M^c)$.

2.11 Example: Gambler's Ruin

You enter a casino with $\$k$, and on each spin of a roulette wheel you bet $\$1$ at evens on the event R that the result is red. The wheel is not fair, so $\mathbf{P}(R) = p < \frac{1}{2}$. If you lose all $\$k$,

you leave; and if you ever possess $\$K \geq \k , you choose to leave immediately. What is the probability that you leave with nothing? (Assume spins of the wheel are independent. Note that this is an idealized casino.)

Solution Let p_k be the probability that you leave with nothing. If the first spin results in red, you gain $\$1$ and are in the same position as if you had just entered with $\$(k + 1)$. Thus, conditional on R , your chance of leaving with nothing is p_{k+1} . Similarly, if the first spin results in black (or zero), you have $\$(k - 1)$ and your chance of leaving with nothing is p_{k-1} . Hence,

$$(1) \quad p_k = pp_{k+1} + (1 - p)p_{k-1}, \quad 0 < k < K.$$

If $k = 0$, then you certainly leave with nothing; if $k = K$, you leave before betting. Hence, $p_0 = 1$ and $p_K = 0$.

Writing (1) as

$$p_{k+1} - p_k = \frac{1 - p}{p}(p_k - p_{k-1}) \quad \text{for } p > 0,$$

gives (on using $p_0 = 1$)

$$p_{k+1} - p_k = \left(\frac{1 - p}{p}\right)^k (p_1 - 1), \quad \text{by Theorem 2.3.1(a).}$$

Hence,

$$p_k = 1 + (p_1 - 1) \frac{\left(\frac{1 - p}{p}\right)^k - 1}{\frac{1 - p}{p} - 1}, \quad \text{by Theorem 2.3.1(b).}$$

Because $p_K = 0$, setting $k = K$ in this gives

$$0 = 1 + (p_1 - 1) \frac{\left(\frac{1 - p}{p}\right)^K - 1}{\frac{1 - p}{p} - 1}.$$

and now eliminating p_1 gives finally

$$(2) \quad p_k = \frac{\left(\frac{1 - p}{p}\right)^K - \left(\frac{1 - p}{p}\right)^k}{\left(\frac{1 - p}{p}\right)^K - 1}.$$

(3) **Exercise** Show that as $K \rightarrow \infty$ in (2), $p_k \rightarrow 1$.

(4) **Exercise** Find p_k when $p = \frac{1}{2}$.

(5) **Exercise** Show that with probability one you do not remain in the casino forever.

(6) **Exercise** Given that you leave with nothing, what is the probability that you never possessed more than your initial $\$(k)$?

(7) **Exercise** Let $n(k, K, r)$ be the number of sequences of red and black that result in your leaving the casino with $\$(k)$ on the r th spin of the wheel. Show that the probability of this event is $n(k, K, r)p^{(r+K-k)/2}(1 - p)^{(r-K+k)/2}$.

- (8) **Exercise** Let $K = 2k$. Show that the probability that you leave the casino on the r th spin given that you leave with nothing is the same as the probability that you leave the casino on the r th spin given that you leave with $\$K$.
- (9) **Exercise** Show that doubling the stakes increases your chance of leaving with $\$K$. What does this tell you about most gamblers?

2.12 Example: Accidents and Insurance

In any given year, the probability that a given male driver has a mishap entailing a claim from his insurance company is μ , independently of other years. The equivalent probability in female drivers is λ .

Assume there are equal numbers of male and female drivers insured with the Acme Assurance Association, which selects one of them at random.

- (a) What is the probability that the selected driver makes a claim this year?
 (b) What is the probability that the selected driver makes a claim in two consecutive years?
 (c) If the insurance company picks a claimant at random, what is the probability that this claimant makes another claim in the following year?

Solution (a) Let A_1 and A_2 be the events that a randomly chosen driver makes a claim in each of the first and second years. Then conditioning on the sex of the driver (M or F) yields

$$\mathbf{P}(A_1) = \mathbf{P}(A_1|M)\mathbf{P}(M) + \mathbf{P}(A_1|F)\mathbf{P}(F) = \frac{1}{2}(\mu + \lambda)$$

because $\mathbf{P}(F) = \mathbf{P}(M) = \frac{1}{2}$.

(b) Likewise,

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1 \cap A_2|M)\mathbf{P}(M) + \mathbf{P}(A_1 \cap A_2|F)\mathbf{P}(F) = \frac{1}{2}(\mu^2 + \lambda^2).$$

(c) By definition,

$$\mathbf{P}(A_2|A_1) = \mathbf{P}(A_2 \cap A_1)/\mathbf{P}(A_1) = \frac{\mu^2 + \lambda^2}{\mu + \lambda}$$

- (1) **Exercise** Note that A_1 and A_2 are conditionally independent given the sex of the driver. Are they ever independent?
- (2) **Exercise** Show that $\mathbf{P}(A_2|A_1) \geq \mathbf{P}(A_2)$. When does equality hold?
- (3) **Exercise**
- (a) Find the probability that a driver makes a claim in a third year given that the driver has claimed in each of the two preceding years.
 (b) Find the probability that a driver claims in year n , given that this driver has claimed in all of the preceding n years.
 (c) Find the limit in (b) as $n \rightarrow \infty$.
- (4) **Exercise** Find the probability that a claimant is
- (a) Male
 (b) Female.
- (5) **Exercise** Find the probability that a driver claiming in n successive years is male.

- (6) **Exercise** Now rework the example and exercises without assuming that equal numbers of male and female drivers are insured with the AAA.

Remark The fact that a claimant is more likely to have a second accident, even though accidents were assumed independent for a given driver, is an example of what is sometimes called a sampling paradox (though it is not a paradox). It is the reason why insurance companies offer no-claims discounts (or at least, one of the reasons). It is the case in practice that $\mu > \lambda$.

2.13 Example: Protocols

Part A: Boys and Girls

Consider the following question: “Tom has exactly one sibling. What is the probability that it is a girl?”

- (a) Do you think this question has a well-defined answer?
 (b) If so, write down your answer, and then consider the following arguments:
- (i) There are three family possibilities; two girls, two boys, or one of each. Two girls is impossible, which leaves equal chances that the sibling is a boy or a girl. The answer is $\frac{1}{2}$.
 - (ii) Families with a child called Tom arise in four equally likely ways: TB, BT, TG, GT . So Tom has a brother as often as he has a sister. The answer is $\frac{1}{2}$.
 - (iii) There are four cases: BB, BG, GB, GG . The last is impossible, and in two of the remaining three cases the sibling is a girl. The answer is $\frac{2}{3}$.
 - (iv) Assuming that the sex of siblings is independent, the other sibling is equally likely to be a girl or a boy. The answer is $\frac{1}{2}$.

Are any of these correct? Is yours correct? (You may assume that any given birth gives rise to one girl or one boy with equal probability.)

Solution (a) The question is ill-posed; there is no correct answer because the sample space is not defined. This is the same as saying that the underlying experiment (selecting Tom) is not described.

(b) We may consider some well-posed questions.

I A woman has two children that are independently equally likely to be a boy or a girl. One of them at least (Tom) is male. Now the sample space has four equally likely outcomes:

$$(1) \quad \Omega = \{BG, GB, BB, GG\};$$

the event of interest is $A = \{BG, GB\}$, and we are given that B occurs, where $B = \{BG, GB, BB\}$. Hence,

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{2}{3}.$$

II A woman has two children that are independently equally likely to be a boy or a girl. Her first son is called Tom with probability $p_1 < 1$. If she has two sons, and the oldest is

not Tom, the second son is called Tom with probability p_2 . The sample space is

$$(2) \quad \Omega = \{BB, BT, TB, TG, BG, GT, GB, GG\}$$

where, for example, $\mathbf{P}(BT) = \frac{1}{4}(1 - p_1)p_2$. Then the required probability is, using Theorem 2.1.4,

$$\frac{\mathbf{P}(GT) + \mathbf{P}(TG)}{\mathbf{P}(TB) + \mathbf{P}(BT) + \mathbf{P}(TG) + \mathbf{P}(GT)} = \frac{p_1 + p_2}{2p_1 + 2p_2 - p_1p_2}.$$

Notice that this is equal to $\frac{1}{2}$ if either $p_1 = 0$ or $p_2 = 0$, but not both. It is also $\frac{1}{2}$ if $p_1 = 1$ and we define p_2 to be zero. In any case,

$$\frac{1}{2} \leq \frac{p_1 + p_2}{2p_1 + 2p_2 - p_1p_2} \leq \frac{2}{3}.$$

Notice that we have assumed that families and names are independent; that is, that women are not more (or less) likely to have boys because they want to call them Tom (or not), and that having a girl does not change the chance that a boy is called Tom.

III A boy is selected at random from a number of boys who have one sibling. This sample space has four equally likely outcomes

$$\Omega = \{B^*B, BB^*, B^*G, GB^*\}$$

where the star denotes the boy (Tom) who was picked at random. (The experiment amounts to picking one of the B -symbols in (1) with equal chance of picking any.) Hence, the event of interest is $A = \{B^*G, GB^*\}$ and the required probability is $\frac{1}{2}$.

IV A chance acquaintance is introduced as Tom who has just one sibling. What is the chance that it is a sister?

The sample space is the set of your chance acquaintances. This is too vague to allow further progress.

Remark The arguments of (b) (i), (ii), (iii), and (iv) appeared in letters to *The Guardian* in June 1989. An answer can only be defined when the exact procedure (also known as a *protocol*) for selecting Tom is decided. If, for example, you meet Tom at a club for identical twins, the problem is different again.

Notice also that parts of this example are getting rather distant from the type of experiment used to justify our axioms. You may well see no particular reason to suppose that our theory of probability is relevant in, say, Case IV, or even in Case II.

(3) **Exercise** In the framework of Case II, consider the following two procedures:

- (a) Select one of her two children at random
- (b) Select one of her sons (if any) at random.

In each case, find the probability that the child is the elder given that his name is Tom.

(4) **Exercise** In the framework of Case II, can it be the case that TG , TB , BT , and GT are equally likely outcomes?

(5) **Exercise** Suppose a woman has three children and each is independently equally likely to be male or female. Show that the event “they are either all girls or all boys” is independent of the event “at least two children are boys.”

Part B: Goats and Cars: The Monty Hall Problem

Suppose yourself to be participating in the following bizarre contest. You have a choice of three doors. Behind one door is a costly automobile, behind the other two doors are cheap goats. You choose the first door, whereupon the master of ceremonies opens the third door to reveal a goat; he then offers you the opportunity to change your choice of door. Can you calculate the probability that the car lies behind the second door? (You are given the object behind the door you open.)

Solution No, you cannot. To see this, let C_i be the event that the car lies behind the i th door, and let G be the event that a goat is revealed to you behind the third door. You require $\mathbf{P}(C_2|G)$, which we can write as

$$(6) \quad \mathbf{P}(C_2 \cap G)/\mathbf{P}(G) = \frac{\mathbf{P}(G|C_2)\mathbf{P}(C_2)}{\mathbf{P}(G|C_1)\mathbf{P}(C_1) + \mathbf{P}(G|C_2)\mathbf{P}(C_2)} = \frac{\mathbf{P}(G|C_2)}{\mathbf{P}(G|C_1) + \mathbf{P}(G|C_2)},$$

on the reasonable assumption that the car is equally likely to be behind any door, so that $\mathbf{P}(C_2) = \mathbf{P}(C_1)$. Now observe that all three terms in the denominator and numerator of (6) depend on the decisions of the master of ceremonies. His rules for making his decision once again form a *protocol*. If you do not know his protocol for the contest, you cannot calculate $\mathbf{P}(C_2|G)$.

Remark This problem was presented in *Parade* magazine (1990, 1991, distributed in the USA) and generated an extensive correspondence in that and several other periodicals. Almost all participants assumed (wrongly) that the problem as stated has one solution, and chiefly disputed as to whether the answer should be $\frac{1}{2}$ or $\frac{2}{3}$.

- (7) **Exercise** Show that if you have paid the master of ceremonies enough to ensure that you win, then $\mathbf{P}(C_2|G) = 1$.
- (8) **Exercise** Show that if the master of ceremonies has decided that (i) whatever you choose, he will show you a goat; and (ii) if he has a choice of two goats, he will pick one at random, then $\mathbf{P}(C_2|G) = \frac{2}{3}$.
- (9) **Exercise** Show that if the master of ceremonies has decided that (i) whatever you choose, he will open a different door; and (ii) he will pick it at random, then $\mathbf{P}(C_2|G) = \frac{1}{2}$.
- (10) **Exercise** Show that if the master of ceremonies has decided that (i) if a goat is behind the first door, he will open it for you; and (ii) if a car lies behind the first door, he will open another door, then $\mathbf{P}(C_2|G) = 0$.

Remark There are many famous problems equivalent to these two, to all of which the correct answer is, *there is no unique answer* (e.g., the ‘Prisoners’ paradox’, ‘Red Ace’). There seems to be no way of preventing the futile, acrimonious, and incorrect discussions accompanying their regular appearance in the popular press. The so-called ‘Doomsday Argument’ provides a slightly different but equally fallacious example.

2.14 Example: Eddington’s Controversy

Four men each tell the truth independently with probability $\frac{1}{3}$. D makes a statement that C reports to B , and B then reports C ’s statement to A . If A asserts that B denies that C claims that D is a liar, what is the probability that D spoke the truth?

Solution Let S_A be the event that A makes the given statement. Further, let A denote the event that A tells the truth, A^c denote the event that A lies, and so on. Then, obviously,

$$D \cap C \cap B \cap A \subseteq S_A,$$

for if they all tell the truth then A makes the given statement. Also, by the independence,

$$\mathbf{P}(D \cap C \cap B \cap A) = \frac{1}{81}.$$

Likewise, by following through the chain of assertions, we see that $D \cap C^c \cap B^c \cap A$, $D \cap C^c \cap B \cap A^c$ and $D \cap C \cap B^c \cap A^c$ are included in S_A , each having probability $\frac{4}{81}$.

When D lies,

$$D^c \cap C^c \cap B^c \cap A^c \subseteq S_A$$

for A also makes the given statement if they are all liars. Here,

$$\mathbf{P}(D^c \cap C^c \cap B^c \cap A^c) = \frac{16}{81},$$

and likewise $D^c \cap C^c \cap B \cap A$, $D^c \cap C \cap B^c \cap A$ and $D^c \cap C \cap B \cap A^c$ are included in S_A , each having probability $\frac{4}{81}$. These mutually exclusive outcomes exhaust the possibilities, so by conditional probability

$$\mathbf{P}(D|S_A) = \frac{\mathbf{P}(D \cap S_A)}{\mathbf{P}(D \cap S_A) + \mathbf{P}(D^c \cap S_A)} = \frac{\frac{13}{81}}{\frac{13}{81} + \frac{28}{81}} = \frac{13}{41}.$$

- (1) **Exercise** What is the probability that C did claim that D is a liar, given S_A ?
- (2) **Exercise** What is the probability that both C and D lied, given S_A ?
- (3) **Exercise** Prove the result of the example more laboriously by using conditional probability, rather than by listing outcomes.
- (4) **Exercise** Eddington himself gave the answer to this problem as $\frac{25}{71}$. This is the controversy! Can you reconstruct the argument that led him to this answer?

Remark This example is similar to Example 2.13, in that it is entertaining but of no real practical value. Our theory of probability does not pretend to include this type of problem, and nothing can be said about the credibility of real reports by these methods. Despite this, the first attempt to do so was made in the seventeenth century, and such attempts have been repeated sporadically ever since.

PROBLEMS

- 1 The probability that an archer hits the target when it is windy is 0.4; when it is not windy, her probability of hitting the target is 0.7. On any shot, the probability of a gust of wind is 0.3. Find the probability that:
 - (a) On a given shot, there is a gust of wind and she hits the target.
 - (b) She hits the target with her first shot.
 - (c) She hits the target exactly once in two shots.
 - (d) There was no gust of wind on an occasion when she missed.

- 2 Let A, B be two events with $\mathbf{P}(B) > 0$. Show that
- If $B \subset A$, then $\mathbf{P}(A|B) = 1$,
 - If $A \subset B$, then $\mathbf{P}(A|B) = \mathbf{P}(A)/\mathbf{P}(B)$.
- 3 Three biased coins C_1, C_2, C_3 lie on a table. Their respective probabilities of falling heads when tossed are $\frac{1}{3}, \frac{2}{3}$, and 1. A coin is picked at random, tossed, and observed to fall heads. Calculate the probability that it is C_k for each $k = 1, 2, 3$.
- Given that a coin has been tossed once and observed to fall heads, calculate the probability that a second throw of the same coin will also produce heads.
- The experiment is begun again with the same three coins. This time the coin selected is tossed twice and observed to fall heads both times. Calculate the probability that it is C_k for each $k = 1, 2, 3$.
- Given that a coin has been tossed twice and observed to fall heads both times, calculate the probability that a third throw of the same coin will also produce heads.
- 4
- An event E is independent of itself. Show that it has probability either 0 or 1.
 - Events A and B are disjoint. Can you say whether they are dependent or independent?
 - Prove that if events A and B are independent then so are the events A^c and B , and the events A^c and B^c .
- 5 Candidates are allowed at most three attempts at a given test. Given $j - 1$ previous failures, the probability that a candidate fails at his j th attempt is p_j . If $p_1 = 0.6$, $p_2 = 0.4$, and $p_3 = 0.75$, find the probability that a candidate:
- Passes at the second attempt:
 - Passes at the third attempt:
 - Passes given that he failed at the first attempt;
 - Passes at the second attempt given that he passes.
- 6 Dick throws a die once. If the upper face shows j , he then throws it a further $j - 1$ times and adds all j scores shown. If this sum is 3, what is the probability that he only threw the die
- Once altogether?
 - Twice altogether?
- 7 A man has five coins in his pocket. Two are double-headed, one is double-tailed, and two are normal. They can be distinguished only by looking at them.
- The man shuts his eyes, chooses 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 upper face is a head. What is the probability that the lower face is a head?
 - He shuts his eyes again, picks up the coin, and tosses it again. What is the probability that the lower face is a head?
 - He opens his eyes and sees that the upper face is a head. What is the probability that the lower face is a head?
- 8 An urn contains four dice, one red, one green, and two blue.
- One is selected at random; what is the probability that it is blue?
 - The first is not replaced, and a second die is removed. What is the chance that it is: (i) blue? or (ii) red?
 - The two dice are thrown. What is the probability that they show the same numbers and are the same colour?
 - Now the two remaining in the urn are tossed. What is the probability that they show the same number and are the same colour, given that the first two did not show the same number and colour?
- 9 A 12-sided die A has 9 green faces and 3 white faces, whereas another 12-sided die B has 3 green faces and 9 white faces. A fair coin is tossed once. If it falls heads, a series of throws is made with die A alone; if it falls tails then only the die B is used.

- (a) Show that the probability that green turns up at the first throw is $\frac{1}{2}$.
- (b) If green turns up at the first throw, what is the probability that die A is being used?
- (c) Given that green turns up at the first two throws, what is the probability that green turns up at the third throw?
- 10** Suppose that any child is male with probability p or female with probability $1 - p$, independently of other children. In a family with four children, let A be the event that there is at most one girl, and B the event that there are children of both sexes. Show that there is a value of p , with $0 < p < \frac{1}{2}$, such that A and B are independent.
- 11** Suppose that parents are equally likely to have (in total) one, two, or three offspring. A girl is selected at random; what is the probability that the family includes no older girl? (Assume that children are independent and equally likely to be male or female.)
- 12** Two roads join Ayton to Beaton, and two further roads join Beaton to the City. Ayton is directly connected to the City by a railway. All four roads and the railway are each independently blocked by snow with probability p . I am at Ayton.
- (a) Find the probability that I can drive to the City.
- (b) Find the probability that I can travel to the City.
- (c) Given that I can travel to the City, what is the probability that the railway is blocked?
- 13** An urn contains b blue and r red balls, which may be withdrawn at random according to one of the following three schemes.
- (a) The balls are removed at random one at a time until all those remaining are of the same colour.
- (b) The balls are removed until a ball differs in colour from its predecessor. This first different ball is replaced in the urn; this process is then continued until the remaining balls are all the same colour.
- (c) The balls are removed one by one and inspected. The first is discarded. Each succeeding ball that is the same colour as its predecessor is replaced, the others are discarded, until the remaining balls are all the same colour.
- In each case, find the probability that the remaining balls are all red.
- 14** Let A_1, A_2, \dots, A_n be independent events. Show that the probability that none of the events A_1, \dots, A_n occur is less than $\exp(-\sum_1^n \mathbf{P}(A_i))$.
- 15** Let A and B be independent events. Show that

$$\max\{\mathbf{P}((A \cup B)^c), \mathbf{P}(A \cap B), \mathbf{P}(A \Delta B)\} \geq \frac{4}{9}.$$

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

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

Find a similar expression for $\mathbf{P}(E|A = \text{tail})$ and hence find $\mathbf{P}(E)$.

- 17** After marking the papers of a certain student, the examiners are unable to decide whether he really understands the subject or is just bluffing. They reckon that the probability that he is a bluffer is p , $0 < p < 1$, and the probability that he understands is $q = (1 - p)$. They therefore give him a *viva voce* consisting of n independent questions, each of which has a probability u of being answered by someone who understands the subject. Unfortunately, there is also a probability b , $0 < b < 1$, that the answer can be guessed by someone who does not understand. Show that the probability that the student understands given that he manages to answer k questions correctly is given by r , where

$$r = \frac{qu^k(1-u)^{n-k}}{qu^k(1-u)^{n-k} + pb^k(1-b)^{n-k}}.$$

Show that if the student gets every single question right and $u > b$, then as n increases the probability that the student really understands tends to 1. How many questions must the student get right to convince the examiners that it is more likely that he understands the subject than that he is bluffing?

- 18 A team of three students Amy, Bella, and Carol answer questions in a quiz. A question is answered by Amy, Bella, or Carol with probability $\frac{1}{2}$, $\frac{1}{3}$, or $\frac{1}{6}$, respectively. The probability of Amy, Bella, or Carol answering a question correctly is $\frac{4}{5}$, $\frac{3}{5}$, or $\frac{3}{5}$, respectively. What is the probability that the team answers a question correctly? Find the probability that Carol answered the question given that the team answered incorrectly.

The team starts the contest with one point and gains (loses) one point for each correct (incorrect) answer. The contest ends when the team's score reaches zero points or 10 points. Find the probability that the team will win the contest by scoring 10 points, and show that this is approximately $\frac{4}{7}$.

- 19 A and B play a sequence of games, in each of which A has a probability p of winning and B has a probability q ($= 1 - p$) of winning. The sequence is won by the first player to achieve a lead of two games. By considering what may happen in the first two games, or otherwise, show that the probability that A wins the sequence is $p^2/(1 - 2pq)$.

If the rules are changed so that the sequence is won by the player who first wins two consecutive games, show that the probability that A wins the sequence becomes $p^2(1 + q)/(1 - pq)$.

Which set of rules gives that weaker player the better chance of winning the sequence?

- 20 You toss a coin. If it shows a tail, you roll one die and your score is the number it shows. If the coin shows a head, you toss five more coins and your score is the total number of heads shown (including the first coin). If you tell me only that your score is two, what is the probability that you rolled a die?
- 21 Three fair dice labelled A, B, and C are rolled on to a sheet of paper. If a pair show the same number a straight line is drawn joining them. Show that the event that the line AB is drawn is independent of the event that BC is drawn. What is the probability that a complete triangle is drawn? (The dice are not colinear.)
- 22 You roll a fair die n times. What is the probability that
- You have rolled an odd number of sixes?
 - You have not rolled a six on two successive rolls?
 - You rolled a one before you rolled a six, given that you have rolled at least one of each?
- 23 Irena throws at a target. After each throw she moves further away so that the probability of a hit is two-thirds of the probability of a hit on the previous throw. The probability of a hit on the first throw is $\frac{1}{4}$. Find the probability of a hit on the n th throw. Deduce that the probability of never hitting the target is greater than $\frac{1}{4}$.
- 24 A fair coin is tossed three times. What is the probability that it lands "heads" at least once?

In a coin-tossing game, a player tosses five fair coins. If he is content with the result, he stops. If not, he picks up one or more of the coins and tosses them a second time. If he is still dissatisfied, he may for one last time pick up and throw again one or more of the coins. Show that if the player's aim is to finish with five heads showing, and if he uses the best strategy, then the probability that he will succeed is $(\frac{7}{8})^5$.

A second player plays the same game but aims to finish with *either* all heads *or* all tails showing. What is the probability of his succeeding?

- 25 Alf and Bert play a game that each wins with probability $\frac{1}{2}$. The winner then plays Charlie whose probability of winning is always θ . The three continue in turn, the winner of each game always playing the next game against the third player, until the tournament is won by the first player to win two successive games. Let p_A , p_B , p_C be the probabilities that Alf, Bert, and Charlie, respectively, win the tournament. Show that $p_C = 2\theta^2/(2 - \theta + \theta^2)$. Find p_A and p_B , and find the value of θ for which p_A , p_B , p_C are all equal. (Games are independent.)

If Alf wins the tournament, what is the probability that he also won the first game?

- 26 Box A contains three red balls and two white balls; box B contains two red balls and two white balls. A fair die is thrown. If the upper face of the die shows 1 or 2, a ball is drawn at random from

box A and put in box B and then a ball is drawn at random from box B . If the upper face of the die shows 3, 4, 5 or 6, a ball is drawn at random from box B and put in box A , and then a ball is drawn at random from box A .

What are the probabilities

- (a) That the second ball drawn is white?
 (b) That both balls drawn are red?
 (c) That the upper face of the red die showed 3, given that one ball drawn is white and the other red?
- 27** A fair six-sided die, with faces numbered from 1 to 6, is thrown repeatedly on to a flat surface until it first lands with the 6 face uppermost. Find the probability that this requires:
- (a) n throws.
 (b) An even number of throws.
 (c) Show that the probability that the 5 face appears at least once before the first 6 is $\frac{1}{2}$, and find the probability that all the faces 1 to 5 appear before the first 6.
- 28** Suppose that n water lily leaves are placed so that the base of each leaf lies on a circle. A frog is initially on leaf L_1 ; she hops clockwise to the adjacent leaf L_2 with probability p , or anticlockwise to leaf L_n with probability q . Succeeding hops are independent, and go to the nearest leaf clockwise with probability p or the nearest leaf anticlockwise with probability q .

Find the probability that:

- (a) The frog returns to L_1 before visiting all n leaves.
 (b) The first hop on to L_1 has the same orientation as the first hop off L_1 .
 (c) What is the probability that the first hop on to L_1 is clockwise?
- 29** Anselm and Bill toss a fair coin repeatedly. Initially, Anselm has m marks, where $1 \leq m \leq n - 1$. If the coin shows a head, then Anselm gains a mark from Bill; otherwise, he forfeits a mark to Bill. Whenever Anselm has n marks, he must immediately give one to Bill.

Let p_m^k be the probability that Anselm has n marks on k occasions before the first moment at which he has no marks. Write down a difference equation for p_m^k , with appropriate boundary conditions, and deduce that for $k \geq 1$,

$$p_m^k = \frac{m}{n^2} \left(1 - \frac{1}{n}\right)^{k-1}.$$

Explain how you could have shown this without solving the equation for p_m^k . Show that Anselm is certain to lose all his marks eventually.

- 30** A and B each have \$60. They play a sequence of independent games at each of which A wins \$ x from B with probability p , or loses \$ x to B with probability q , where $p + q = 1$. The stake x is determined by rolling a fair die once, and setting x as the number shown by the die; $1 \leq x \leq 6$.
- (a) What is the probability that A wins his opponent's fortune before losing his own?
 (b) If A could choose the stake to be an integer x such that $1 \leq x \leq 6$, and $p < q$, what value should he choose for x ?
- 31** A document is equally likely to be in any of three boxfiles. A search of the i th box will discover the document (if it is indeed there) with probability p_i .

What is the probability that the document is in the first box:

- (a) Given that I have searched the first box once and not found it?
 (b) Given that I have searched the first box twice and not found it?
 (c) Given that I have searched all three boxes once and not found it?
- Assume searches are independent.
- 32** A network forming the edges of a cube is constructed using 12 wires, each 1 metre long. An ant is placed on one corner and walks around the network, leaving a trail of scent as it does so. It never turns around in the middle of an edge, and when it reaches a corner:

- (i) If it has previously walked along both the other edges, it returns along the edge on which it has just come.
- (ii) If it has previously walked along just one of the other edges, it continues along the edge along which it has not previously walked.
- (iii) Otherwise, it chooses one of the other edges arbitrarily.

Show that the probability that the ant passes through the corner opposite where it started after walking along just three edges is $\frac{1}{2}$, but that it is possible that it never reaches the opposite corner. In the latter case, determine the probability of this occurring. What is the greatest distance that the ant has to walk before an outside observer (who knows the rules) will know whether the ant will ever reach the corner opposite where it started?

Show that the rules may be modified to guarantee that the ant (whose only sense is smell) will be able to reach the corner opposite the corner where it started by walking not more than a certain maximum distance that should be determined. The ant can count.

- 33 Pooling** You have $(n!)^2$ jars of fluid, one of which is contaminated. Any jar is equally likely to be the contaminated one, and you need to identify it. Luckily, you have an infallible test.
- (a) If you test the jars one at a time, find the probability that you require t tests to identify the contaminated jar.
 - (b) Alternatively, you may arrange the jars in j groups of size k , where $jk = (n!)^2$. A sample from each of the jars in a group is pooled in one jar, and this pooled sample is tested. On finding the contaminated pooled sample, each jar of this group is tested separately. Find the probability that you require t tests to find the contaminated jar.
 - (c) What is the best choice for j and k ?
- 34 Simpson's Paradox** Two drugs are being tested. Of 200 patients given drug A , 60 are cured; and of 1100 given drug B , 170 are cured. If we assume a homogeneous group of patients, find the probabilities of successful treatment with A or B . Now closer investigation reveals that the 200 patients given drug A were in fact 100 men, of whom 50 were cured, and 100 women of whom 10 were cured. Further, of the 1100 given drug B , 100 were men of whom 60 were cured, and 1000 were women of whom 110 were cured. Calculate the probability of cure for men and women receiving each drug; note that B now seems better than A . (Results of this kind indicate how much care is needed in the design of experiments. Note that the paradox was described by Yule in 1903, and is also called the **Yule-Simpson paradox**.)
- 35** In Problem 34, given that a randomly chosen patient is cured, find:
- (a) The probability that the patient is male.
 - (b) The probability that the patient is female.
- 36 Prisoners' Paradox** Three prisoners are informed by their warder that one of them is to be released and the other two shipped to Devil's Island, but the warder cannot inform any prisoner of that prisoner's fate. Prisoner A thus knows his chance of release to be $\frac{1}{3}$. He asks the warder to name some one of the other two who is destined for Devil's Island, and the warder names B . Can A now calculate the conditional probability of his release?
- 37** Let A and B be events. Show that $\mathbf{P}(A \cap B | A \cup B) \leq \mathbf{P}(A \cap B | A)$. When does equality hold?
- 38** Explain the following "paradox" posed by Lewis Carroll. We are provided with a supply of balls that are independently equally likely to be black or white.

Proposition *If an urn contains two such balls, then one is black and the other white.*

Proof Initially, $\mathbf{P}(BB) = \mathbf{P}(BW) = \mathbf{P}(WB) = \mathbf{P}(WW) = \frac{1}{4}$. Add a black ball, so that now

$$\mathbf{P}(BBB) = \mathbf{P}(BBW) = \mathbf{P}(BWB) = \mathbf{P}(BWW) = \frac{1}{4}.$$

Now pick a ball at random. By (2.1.3),

$$\mathbf{P}(\text{black ball drawn}) = 1 \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} + \frac{2}{3} \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{4} = \frac{2}{3}.$$

But if I pick a ball at random from three, with probability $\frac{2}{3}$ of drawing a black ball, then two are black and one is white. Hence, before adding the black ball the urn contained one white ball and one black ball. ■

- 39** Let M_1, M_2, \dots, M_n be a sequence of men such that M_j reports to M_{j+1} on a statement made by M_{j-1} . Let R_n be the event that M_n reports that M_{n-1} reports that \dots that M_2 reports that M_1 is a liar. If each reporter lies independently with probability p , find p_n , the probability that M_1 told the truth given R_n . Show that as $n \rightarrow \infty$, $p_n \rightarrow 1 - p$.

- 40** Suppose that for events S , A , and B ,

$$\begin{aligned} (*) \quad & \mathbf{P}(S|A) \geq \mathbf{P}(S) \\ & \mathbf{P}(A|S \cap B) \geq \mathbf{P}(A|S) \\ & \mathbf{P}(A|S^c) \geq \mathbf{P}(A|S^c \cap B). \end{aligned}$$

(a) Show that, except in trivial cases, $\mathbf{P}(S|A \cap B) \geq \mathbf{P}(S|B)$.

(b) Show that $\mathbf{P}(S|A) \geq \mathbf{P}(A)$.

(c) Show that if (*) is replaced by $\mathbf{P}(S|B) \geq \mathbf{P}(S)$, then $\mathbf{P}(S|A \cap B) \geq \mathbf{P}(S|A)$.

- 41** You have to play Alekhine, Botvinnik, and Capablanca once each. You win each game with respective probabilities p_a , p_b , and p_c , where $p_a > p_b > p_c$. You win the tournament if you win two consecutive games, otherwise you lose, but you can choose in which order to play the three games. Show that to maximize your chance of winning you should play Alekhine second.

- 42** Show that the events A and B are mutually attractive if and only if $\mathbf{P}(B|A) > \mathbf{P}(B|A^c)$.

- 43 Weather** Days can be sunny or cloudy. The weather tomorrow is the same as the weather today with probability p , or it is different with probability q , where $p + q = 1$. If it is sunny today, show that the probability s_n that it will be sunny n days from today satisfies

$$s_n = (p - q)s_{n-1} + q; \quad n \geq 1,$$

where $s_0 = 1$. Deduce that

$$s_n = \frac{1}{2}(1 + (p - q)^n); \quad n \geq 1.$$

- 44 Flats, Sharps, and Craps** Dice can be crooked (or weighted) in various ways. One way is to shorten the distance between one pair of opposite faces, thus making them more likely; these are called *flats* (also known as *broads* or *doctors*).

Another way is to taper all four sides of the die, creating a truncated pyramid, or to insert a weight in the base. The top face becomes more likely and the base less likely; these are called *sharps*.

You have three pairs of dice:

- (i) A fair pair for which $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$.
 - (ii) A pair of 1–6 flats for which $p_1 = p_6 = \frac{1}{4}$ and $p_2 = p_3 = p_4 = p_5 = \frac{1}{8}$.
 - (iii) A pair of 5–2 sharps for which $p_5 = \frac{1}{4}$, $p_2 = \frac{1}{12}$, and $p_1 = p_3 = p_4 = p_6 = \frac{1}{6}$.
- With which pair would you prefer to play craps? (See Example 1.12 for the rules.)

- 45 The Monty Hall Problem: Example 2.13B** Suppose that the presenter's protocol requires him to show you a goat when he opens another door. With a choice of two goats (called Bill and Nan, say), he shows you Bill with probability b . Show that the conditional probability that the third door conceals the car, given that you are shown Bill, is $\frac{1}{1+b}$.

3

Counting

What I say is, patience, and shuffle the cards.
Cervantes

This chapter deals with a special subject and may be omitted on a first reading. Its contents are important and useful, but are not a prerequisite for most of the following chapters.

3.1 First Principles

We have seen that many interesting problems in probability can be solved by counting the number of outcomes in an event. Such counting often turns out to also be useful in more general contexts. This chapter sets out some simple methods of dealing with the commonest counting problems.

The basic principles are pleasingly easy and are perfectly illustrated in the following examples.

- (1) **Principle** If I have m garden forks and n fish forks, then I have $m + n$ forks altogether.
- (2) **Principle** If I have m different knives and n different forks, then there are mn distinct ways of taking a knife and fork.

These principles can be rephrased in general terms involving objects, operations, or symbols and their properties, but the idea is already obvious. The important points are that in (1), the two sets in question are disjoint; that is a fork cannot be both a garden fork and a fish fork. In (2), my choice of knife in no way alters my freedom to choose any fork (and vice versa).

Real problems involve, for example, catching different varieties of fish, drawing various balls from a number of urns, and dealing hands at numerous types of card games. In the standard terminology for such problems, we say that a number n (say) of *objects* or *things* are to be *divided* or *distributed* into r *classes* or *groups*.

The number of ways in which this distribution can take place depends on whether

- (i) The objects can be distinguished or not.
- (ii) The classes can be distinguished or not.
- (iii) The order of objects in a class is relevant or not.

- (iv) The order of classes is relevant or not.
- (v) The objects can be used more than once or not at all.
- (vi) Empty classes are allowed or not.

We generally consider only the cases having applications in probability problems. Other aspects are explored in books devoted to combinatorial theory.

- (3) **Example** (a) Six dice are rolled. What is the probability that they all show different faces?
 (b) What is the probability that five dice show different faces when rolled?

Solution (a) Let A be the event that they all show different faces. Then, because of the assumed symmetry of the dice,

$$\mathbf{P}(A) = \frac{|A|}{|\Omega|}.$$

Now the upper face of each die may be freely chosen in six different ways, so by Principle 2

$$|\Omega| = 6^6 = 46656.$$

However, for outcomes in A , the upper faces are required to be different. Thus, when the upper face of one die is freely chosen in six ways, the upper face of the next can be freely chosen in five ways (different from the first choice). The next may be freely chosen in four ways, and so on. Hence, by Principle 2, $|A| = 6!$ and

$$\mathbf{P}(A) = \frac{6!}{6^6} = \frac{5}{324}.$$

(b) Let \hat{A} be the event that the five dice show different faces. By the same argument as above, $|\Omega| = 6^5$.

For outcomes in \hat{A} we may first make a free choice of which different five faces are to be shown; this is the same as choosing one face not to be shown, which we can do in six ways. Then the first face is freely chosen in five ways, the second in four ways, and so on. Hence, $|\hat{A}| = 6!$ and

$$\mathbf{P}(\hat{A}) = \frac{6!}{6^5} = \frac{5}{54}. \quad \bullet$$

3.2 Permutations: Ordered Selection

Suppose that several objects are placed randomly in a row; playing cards or lottery numbers provide trite examples (but important if you hold a ticket). The number of ways in which this arrangement may occur depends on how many objects there are, whether they are all distinct, whether they may be repeated, and so on. Such arrangements are called *permutations*.

- (1) **Theorem** Given n distinct symbols, the number of distinct permutations (without repetition) of length $r \leq n$ is

$$n(n-1)\dots(n-r+1) = \frac{n!}{(n-r)!}.$$

- (2) **Theorem** Given n distinct symbols which may be repeated any number of times, the number of permutations of length r is n^r .

Proof Theorems (1) and (2) are easily proved by induction. You do it. ■

- (3) **Theorem** Given $n = \sum_{i=1}^r n_i$ symbols of r distinct types, where n_i are of type i and are otherwise indistinguishable, the number of permutations (without repetition) of all n symbols is

$$M_n(n_1, \dots, n_r) = \frac{n!}{\prod_{i=1}^r n_i!}.$$

Proof Suppose that the symbols of each type are numbered so that they are all distinguishable. Then in each originally unnumbered permutation, the symbols of type 1 can be permuted in $n_1!$ ways, the symbols of type 2 in $n_2!$ ways, and so on. Thus, the total number of permutations is $M_n(n_1, \dots, n_r)n_1!n_2!\dots n_r!$. However, we already know from (1) that the number of permutations of n objects is $n!$. Hence,

$$M_n(n_1, \dots, n_r)n_1!\dots n_r! = n!,$$

which proves (3). ■

The number M_n is known as a *multinomial coefficient*. A particularly important case that arises frequently is when $r = 2$. This is a *binomial coefficient*, and it has its own special notation:

$$(4) \quad M_n(k, n-k) = \binom{n}{k} \quad \text{in most books.}$$

$$= {}^n C_k \quad \text{in some older books.}$$

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

- (5) **Example** You are playing bridge. When you pick up your hand, you notice that the suits are already grouped; that is, the clubs are all adjacent to each other, the hearts likewise, and so on. Given that your hand contains four spades, four hearts, three diamonds, and two clubs, what is the probability $\mathbf{P}(G)$ of this event G ?

Solution There are $13!$ permutations of your hand, which we assume are equally likely by symmetry. Now there are $4!$ permutations in which the spades are adjacent in any given position, $4!$ where the hearts are adjacent, and so on. Furthermore, there are $4!$ permutations of the order in which the respective suits may be placed in their adjacent blocks. Hence, the number of permutations in which G occurs is $4!4!4!3!2!$ and

$$(6) \quad \mathbf{P}(G) = \frac{(4!)^3 3! 2!}{13!} = \frac{4!}{M_{13}(4, 4, 4, 3, 2)}.$$

Alternatively, you may observe from (3) that there are $M_{13}(4, 4, 3, 2)$ permutations of your hand where cards of the same suit are regarded as indistinguishable. For each order of suits, only one of these is in G . Because there are $4!$ permutations of the suit order, we immediately recover (6) again. ●

Finally, we remark that the definition of the symbol $\binom{n}{r}$ is sometimes extended to cases other than that when n and r are integers with $0 \leq r \leq n$.

(7) **Definition** For real x and nonnegative integer r ,

$$\binom{x}{r} = \frac{x(x-1)\dots(x-r+1)}{r!}$$



This definition can occasionally provide more compact expressions, for example, we have

$$\binom{-x}{r} = (-1)^r \binom{x+r-1}{r}.$$

3.3 Combinations: Unordered Selection

In a bridge hand or an election, the order in which you get your cards or the politician his votes is irrelevant. In problems of this type we do not arrange, we choose; a choice of objects or symbols is also called a *combination*.

(1) **Theorem** The number of ways of choosing a set of r symbols from a set of n distinct symbols without repetition is

$$\frac{n!}{r!(n-r)!} = \binom{n}{r}.$$

Proof This is just a special case of Theorem 3.2.3. ■

(2) **Theorem** The number of ways of dividing n distinct objects into r distinct groups of sizes n_1, n_2, \dots, n_r , where $\sum_{i=1}^r n_i = n$, is $n! / \prod_{i=1}^r n_i!$.

Proof This is also a simple corollary of Theorem 3.2.3. ■

(3) **Theorem** Out of $n = \sum_1^r n_i$ symbols (where the i th distinct set contains n_i indistinguishable symbols), we can select $\prod_{i=1}^r (n_i + 1) - 1$ combinations.

Proof Note that we can select any number of symbols from zero to n_i , from each of the r sets, but we cannot take zero from all of them. ■

(4) **Theorem** Given a set of n distinct symbols that may be repeated any number of times, the number of ways of choosing a set of size r is $\binom{n+r-1}{r}$.

Proof A proof of Theorem (4) may be found in Example 3.12 or Theorem 3.7.5. ■

- (5) **Example: Ark** The wyvern is an endangered species in the wild. You want to form a captive breeding colony, and you estimate that a viable colony should initially contain r males and r females. You therefore trap a sequence of animals, each of which is independently male with probability p or female with probability $q = 1 - p$, where $p \neq q$ (the females are more wary). Find the probability p_n that it is necessary to capture n animals to create your viable colony of r males and r females.

Solution Let A_n be the event that you first possess r of each sex with the n th capture, and let M be the event that the n th animal is male. Then, of the previous $n - 1$ captured animals, $r - 1$ are male and $n - r$ are female. For any fixed order of these sexes, the probability of being captured in that order is $p^r q^{n-r}$. The number of ways of ordering r males and $n - r$ females, with a male last, is just the same as the number of ways of choosing $r - 1$ of the first $n - 1$ captures to be male. By (1), this is $\binom{n-1}{r-1}$. Hence, for $n \geq 2r$,

$$\mathbf{P}(A_n \cap M) = \binom{n-1}{r-1} p^r q^{n-r}.$$

Likewise, when the last animal is female,

$$\mathbf{P}(A_n \cap M^c) = \binom{n-1}{r-1} q^r p^{n-r},$$

and so

$$p_n = \binom{n-1}{r-1} q^r p^r (p^{n-2r} + q^{n-2r}), \quad n \geq 2r. \quad \bullet$$

3.4 Inclusion–Exclusion

If a group of N men contains $N(b_1)$ who are bald, $N(b_2)$ who are bearded, and $N(b_1, b_2)$ who are bald and bearded, how many altogether are bald, bearded, or both? The answer is

$$(1) \quad N_1 = N(b_1) + N(b_2) - N(b_1, b_2),$$

because anyone who is both is counted once in all three terms on the right-hand side, and so contributes just one to the total N_1 in (1).

More generally, if a group of N objects may each have up to r distinct properties b_1, \dots, b_r , then the number possessing at least one is

$$(2) \quad N_1 = \sum_{b_i} N(b_i) - \sum_{b_i < b_j} N(b_i, b_j) + \dots + (-)^{r-1} N(b_1, \dots, b_r).$$

This is proved either by induction or by noting that any object having exactly k of the properties is counted $\binom{k}{t}$ times in the t th term, and it is the case that

$$\sum_{t=1}^k (-)^{t+1} \binom{k}{t} = 1.$$

[See Problem 3.28(a).] Hence, this object contributes just one to the total N_1 , as required. Notice that N_0 , the number of objects possessing none of the r properties, is just

$$(3) \quad N_0 = N - N_1.$$

- (4) **Example: Derangements** A permutation of the first n integers is called a *derangement* if no integer is in its natural position. Thus, $(3, 2, 1)$ is not a derangement of $(1, 2, 3)$, but $(2, 3, 1)$ is a derangement. Suppose one of the $n!$ permutations of $(1, \dots, n)$ is picked at random. Find p_n , the probability that it is a derangement, and show that as $n \rightarrow \infty$, $p_n \rightarrow e^{-1}$.

Solution Let b_k be the property that the integer k is in its natural position (the k th place). Then, the number of derangements is $n! - N_1$, where N_1 is given by (2). Now, $N(b_k)$ is the number of permutations of $(1, \dots, n)$ with k fixed, namely, $(n-1)!$. Likewise, $N(b_i, b_j)$ is the number of permutations of $(1, 2, \dots, n)$ with i and j fixed, namely, $(n-2)!$, and so on. Hence,

$$\begin{aligned} N_1 &= \sum_{j=1}^n (n-1)! - \sum_{1 \leq i < j \leq n} (n-2)! + \dots + (-1)^{n-1} \\ &= n(n-1)! - \binom{n}{2} (n-2)! + \dots + (-1)^{n-1}. \end{aligned}$$

Now, because all permutations were supposed equally likely, the required probability is

$$p_n = \frac{1}{n!}(n! - N_1) = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \dots + (-1)^n \frac{1}{n!} \rightarrow e^{-1}$$

as $n \rightarrow \infty$. ●

3.5 Recurrence Relations

The answer a to a counting problem usually depends on some given parameters; such as the original number n of objects, the number r which are red (say), or the number s selected. We can make this explicit by writing, for example, $a = a(n, r, s)$. It is often possible to find relationships between two or more of $a(n, r, s)$, and then the problem is reduced to solving a recurrence relation. Some of these can indeed be solved.

- (1) **Theorem** If $(a_n; n \geq 0)$ satisfies

$$a_n = c(n)a_{n-1}, \quad n \geq 1,$$

then

$$a_n = \prod_{k=1}^n c(k)a_0.$$

Proof Trivial, by induction. ■

- (2) **Theorem** If $(a_n; n \geq 0)$ satisfies

$$a_{n+2} + 2ba_{n+1} + ca_n = 0, \quad n \geq 0,$$

and $x = \alpha$ and $x = \beta$ are two distinct solutions of $x^2 + 2bx + c = 0$, then

$$a_n = \frac{1}{\alpha - \beta}((a_1 - \beta a_0)\alpha^n - (a_1 - \alpha a_0)\beta^n).$$

Proof Straightforward by induction. ■

Higher-order linear difference equations can be solved similarly; we omit the details. The recurrence can be in more than one variable, as the following example shows.

- (3) **Example** Let $a(n, k)$ be the number of ways of choosing k objects from n objects. Show that

$$a(n, k) = \binom{n}{k}.$$

Solution Suppose we add an additional $(n + 1)$ th object, and choose k of these $n + 1$. This may be done in $a(n + 1, k)$ ways. But also we may consider whether the additional object was among those chosen. The number of choices when it is *not* is $a(n, k)$. If it *is*, then it is necessary to choose $k - 1$ further objects from the remaining n , which we may do in $a(n, k - 1)$ ways. Then, $a(n, k)$ satisfies the difference equation

(4)
$$a(n + 1, k) = a(n, k) + a(n, k - 1), \quad 0 \leq k \leq n,$$

with solution by inspection $a(n, k) = \binom{n}{k}$. ●

The array generated by $a(n, k); 0 \leq k \leq n$, as n increases, is called *Pascal's triangle*; it has many curious and interesting properties, which we do not explore here. Part of it is displayed in Figure 3.1.

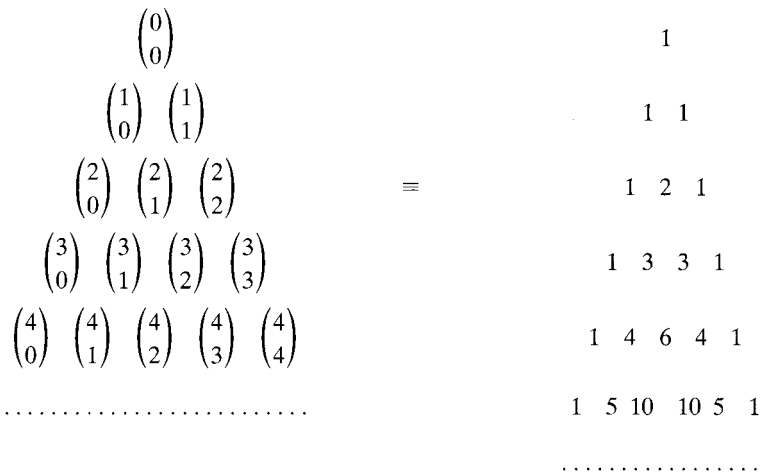


Figure 3.1 Pascal's triangle written as an array of binomial coefficients and as the array of their values. Observe that by (4), each term is obtained by adding its two neighbours in the row above [except for $\binom{0}{0}$ which is defined to be 1].

3.6 Generating Functions

Even quite straightforward counting problems can lead to laborious and lengthy calculations. These are often greatly simplified by using generating functions (introduced by de Moivre and Euler in the early eighteenth century). Later examples will show the utility of generating functions; in this section, we give a fairly bald list of basic definitions and properties, for ease of reference.

- (1) **Definition** Given a collection of numbers $(a_i; i \geq 0)$, the function

$$g_a(x) = \sum_{i=0}^{\infty} a_i x^i$$

is called the generating function of (a_i) . (It is, of course, necessary that $\sum a_i x^i$ converges somewhere if g_a is defined as a function of x . If we regard g_a as an element of a ring of polynomials, such convergence is not necessary.) ▲

- (2) **Definition** Given $(a_i; i \geq 0)$, the function

$$h_a(x) = \sum_{i=0}^{\infty} \frac{a_i x^i}{i!}$$

is the exponential generating function of (a_i) . ▲

- (3) **Definition** Given a collection of functions $(f_n(y); n \geq 0)$; the function

$$g(x, y) = \sum_{n=0}^{\infty} x^n f_n(y)$$

is a bivariate generating function. ▲

The following crucial result is a corollary of Taylor's Theorem. We omit the proof.

- (4) **Theorem (Uniqueness)** If for some x_0 and x_1 we have

$$g_a(x) = g_b(x) < \infty \quad \text{for } x_0 < x < x_1,$$

then $a_i = b_i$ for all i .

Generating functions help to tackle difference equations; the following result is typical.

- (5) **Theorem** If $(a_n; n \geq 0)$ satisfies a recurrence relation

$$a_{n+2} + 2ba_{n+1} + ca_n = d_n; \quad n \geq 0,$$

then

$$g_a(x) = \frac{x^2 g_d(x) + a_0 + a_1 x + 2ba_0 x}{1 + 2bx + cx^2},$$

with a corresponding result for higher-order equations.

Proof To prove (5), multiply each side of the recurrence by x^{n+2} and sum over n . ■

The next four theorems are proved by rearranging the summation on the right-hand side in each case. You should do at least one as an exercise.

- (6) **Theorem (Convolution)** If (a_n) can be written as a convolution of the sequences (b_n) and (c_n) , so

$$a_n = \sum_{i=0}^n c_i b_{n-i}, \quad n \geq 0,$$

then

$$g_a(x) = g_c(x)g_b(x).$$

- (7) **Theorem (Tails)** If

$$b_n = \sum_{i=1}^{\infty} a_{n+i}, \quad n \geq 0,$$

then

$$g_b(x) = \frac{g_a(1) - g_a(x)}{1 - x}.$$

- (8) **Theorem** If

$$c_n = \sum_{i=0}^n a_i,$$

then

$$g_c(x) = \frac{g_a(x)}{1 - x}.$$

- (9) **Theorem (Exponential Function)** Let the function $e(x)$ be defined by

$$e(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!},$$

then

$$e(x + y) = e(x)e(y) = e^{x+y} = e^x e^y.$$

Finally, we have the celebrated

- (10) **Binomial Theorems** For integral n ,

$$(11) \quad (1 + x)^n = \sum_{k=0}^n \binom{n}{k} x^k, \quad n \geq 0$$

and

$$(12) \quad (1 - x)^{-n} = \sum_{k=0}^{\infty} \binom{n+k-1}{k} x^k \quad n \geq 0, |x| < 1.$$

(13) Example Let us prove the binomial theorems.

Proof of (11) Considering the product $(1+x)^n = (1+x)(1+x)\dots(1+x)$, we see that a term x^k is obtained by taking x from any choice of k of the brackets and taking 1 from the rest. Because there are $\binom{n}{k}$ ways of choosing k brackets, the term x^k occurs $\binom{n}{k}$ times in the expansion. Because this is true for any k , the result follows.

Alternatively, you can prove (11) by induction on n , using the easily verified identity

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Proof of (12) Multiplication verifies that $1-x^{m+1} = (1-x)(1+x+\dots+x^m)$. Hence, for $|x| < 1$, we let $m \rightarrow \infty$ to find that (12) is true for $n = 1$,

$$\frac{1}{1-x} = \sum_0^{\infty} x^k.$$

Now we can prove (12) for arbitrary n in a number of ways; one possibility is by induction on n . Alternatively, we observe that a term x^k is obtained in the product $(1+x+x^2+\dots)^n$ if we choose a set of k x s from the n brackets, where we may take any number of x s from each bracket. But by Theorem 3.3.4, this may be done in just $\binom{n+k-1}{k}$ ways. The negative binomial theorem (12) follows. ●

Here is a classic example.

(14) Example: The Coupon Collector's Problem Each packet of an injurious product is equally likely to contain any one of n different types of coupon. If you buy r packets, what is the probability $p(n, r)$ that you obtain at least one of each type of coupon?

Solution This famous problem can be approached in many different ways, as we see later, but the exponential generating function offers a particularly elegant answer.

First, recall from Theorem 3.2.3 that the probability of getting $r = \sum_1^n t_i$ coupons of n distinct types, where t_i are of type i , is

$$n^{-r} M_r(t_1, \dots, t_n) = \frac{r!}{n^r t_1! \dots t_n!}.$$

Then, $p(n, r)$ is the sum of all such expressions in which $t_i \geq 1$ for all i (so we have at least one coupon of each type).

But now expanding $(\sum_{t=1}^{\infty} (s/n)^t / t!)^n$ by the multinomial theorem shows that the coefficient of s^r in this expansion is just $p(n, r)/r!$. Hence,

$$\left(\exp\left(\frac{s}{n}\right) - 1 \right)^n = \sum_{r=n}^{\infty} s^r p(n, r)/r!,$$

and we have obtained the exponential generating function of the $p(n, r)$. ●

Suppose that you are a more demanding collector who requires two complete sets of coupons, let $p_2(n, r)$ be the probability that r packets yield two complete sets of n coupons.

Then exactly similar arguments show that

$$\left(\exp\left(\frac{s}{n}\right) - 1 - \frac{s}{n}\right)^n = \sum_{r=2n}^{\infty} \frac{s^r}{r!} p_2(n, r),$$

and so on for more sets.

In conclusion, it is worth remarking that multivariate generating functions are often useful, although they will not appear much at this early stage. We give one example.

(15) **Multinomial Theorem** Recall the multinomial coefficients defined in Theorem 3.2.3,

$$M_n(n_1, \dots, n_r) = \frac{n!}{\prod_{i=1}^r n_i!}; \quad \sum_{i=1}^r n_i = n.$$

We have

$$(x_1 + x_2 + \dots + x_r)^n = \sum M_n(n_1, \dots, n_r) x_1^{n_1} x_2^{n_2} \dots x_r^{n_r},$$

where the sum is over all (n_1, \dots, n_r) such that $\sum n_i = n$.

Proof This is immediate from the definition of M_n as the number of ways of permuting n symbols of which n_i are of type i . In this case x_i is of type i , of course. ■

Corollary Setting $x_1 = x_2 = \dots = x_r = 1$ gives $\sum M_n(n_1, \dots, n_r) = r^n$.

3.7 Techniques

When evaluating any probability by counting, it is first essential to be clear what the sample space is, and exactly which outcomes are in the event of interest. Neglect of this obvious but essential step has led many a student into lengthy but nugatory calculations.

Second, it is even more important than usual to be flexible and imaginative in your approach. As the following examples show, a simple reinterpretation or reformulation can turn a hard problem into a trivial one. The main mechanical methods are:

- (i) Using the theorems giving the numbers of ordered and unordered selections, recalling in particular that the number of ordered selections of r objects is the number of unordered selections multiplied by $r!$
- (ii) Use of the inclusion–exclusion principle.
- (iii) Setting up recurrence relations.
- (iv) Use of generating functions.

Third, we remark that the implicit assumption of this chapter can be turned on its head. That is, we have developed counting techniques to solve probability problems, but the solution of probability problems can just as well be used to prove combinatorial identities.

(1) **Example** Prove the following remarkable identity:

$$(2) \quad \sum_{k=0}^n \binom{n+k}{k} 2^{-(n+k)} = 1.$$

Hint: Consider an ant walking on a square lattice.

Solution An ant walks on the square lattice of points with nonnegative integer coordinates (i, j) , $i \geq 0$, $j \geq 0$. It starts at $(0, 0)$. If it is at (x, y) , it proceeds next either to $(x + 1, y)$ or to $(x, y + 1)$ with equal probability $\frac{1}{2}$. Therefore, at some transition (certainly less than $2n + 1$ transitions), it leaves the square

$$(0 \leq x \leq n, 0 \leq y \leq n).$$

It does so either with an x -step from (n, y) to $(n + 1, y)$, where $0 \leq y \leq n$ or with a y -step from (x, n) to $(x, n + 1)$, where $0 \leq x \leq n$. Let these $2n + 2$ events be S_y ($0 \leq y \leq n$) and S_x ($0 \leq x \leq n$), respectively. Then by symmetry

$$(3) \quad \sum_{x=0}^n \mathbf{P}(S_x) = \sum_{y=0}^n \mathbf{P}(S_y) = \frac{1}{2}.$$

However, S_y occurs if the ant has taken exactly y vertical steps before its $(n + 1)$ th horizontal step. There are $\binom{n+y+1}{y}$ choices for the y vertical steps and each route to $(n + 1, y)$ has probability $2^{-(n+y+1)}$. Hence,

$$(4) \quad \mathbf{P}(S_y) = \binom{n+y+1}{y} 2^{-(n+y+1)}.$$

Substituting (4) into (3) yields (2). ●

Finally, we remind the reader of our remark in 1.6 that many counting problems are traditionally formulated as “urn models.” They are used in this context for two reasons; the first is utility. Using urns (instead of some realistic model) allows the student to see the probabilistic features without prejudice from false intuition. The second reason is traditional; urns have been used since at least the seventeenth century for lotteries and voting. (Indeed the French expression “aller aux urnes” means “to vote”.) It was therefore natural for probabilists to use them in constructing theoretical models of chance events. A typical example arises when we look at the distribution of n accidents among days of the week, and seek the probability of at least one accident every day or the probability of accident-free days. The problem is equivalent to placing n balls in 7 urns, and seeking the number of ways in which the outcome of interest occurs.

The following result is central here.

(5) **Theorem** Let (x_1, \dots, x_n) be a collection of integers, such that $x_1 + \dots + x_n = r$. The number of distinct ways of choosing them is:

$$(a) \quad \binom{n+r-1}{n-1} \text{ if } x_i \geq 0 \text{ for all } i.$$

$$(b) \quad \binom{r-1}{n-1} \text{ if } x_i > 0 \text{ for all } i.$$

Note that (a) is equivalent to the number of ways of placing r indistinguishable balls in n numbered urns, and (b) adds the additional requirement that there must be at least one ball in each urn. Also, (a) is equivalent to Theorem 3.3.4.

Proof Using the urn model formulation, we simply imagine the r balls placed in a line. Then allocate them to urns by placing $r - 1$ dividers in the $n - 1$ gaps between the balls. Each such choice supplies a distinct set (x_1, \dots, x_n) with $x_i > 0$, and there are $\binom{r-1}{n-1}$ such choices, by (3.3.1). This proves (b). For (a), we add n balls to the group, allocate them in $\binom{n+r-1}{n-1}$ ways by (b), and then remove one ball from each urn. All such allocations form a one-one correspondence with an allocation with $x_i \geq 0$ and $\sum_1^n x_i = r$. This proves (a). ■

3.8 Review and Checklist for Chapter 3

Many chance experiments have equally likely outcomes so that probabilities may be evaluated by counting outcomes in events of interest. Counting can be useful in other contexts also. In this chapter, we introduce a number of techniques for counting, foremost among which is the use of generating functions.

We also introduced inclusion and exclusion, among other methods for counting the outcomes in some event of interest, and many examples to show how these techniques are applied to probability problems. In particular, we have that:

- The number of possible sequences of length r using elements from a set of size n is n^r (with repetition permitted).
- The number of permutations of length r using elements from a set of size $n \geq r$ is $n(n-1) \dots (n-r+1)$ (with repetition not permitted).
- The number of choices (combinations) of r elements from a set of size $n \geq r$ is

$$\binom{n}{r} = \frac{n(n-1) \dots (n-r+1)}{r(r-1) \dots 2 \cdot 1}.$$

- The number of subsets of a set of size n is 2^n .
- The number of permutations of n symbols, of r types, with n_i of the i th type, ($n_1 + \dots + n_r = n$), is

$$M_n(n_1, \dots, n_r) = \frac{n!}{\prod_{i=1}^r n_i!}.$$

- Given a collection of numbers $(a_i; i \geq 0)$ their generating function is $g(x) = \sum_{i=0}^{\infty} a_i x^i$, and their exponential generating function is $h(x) = \sum_{i=0}^{\infty} \frac{a_i x^i}{i!}$.
- The binomial theorems assert that for positive integer n ,

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

and

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

- The multinomial theorem asserts that if $\sum_{i=1}^r n_i = n$, where all quantities are positive integers,

$$(x_1 + x_2 + \cdots + x_r)^n = \sum \frac{n!}{\prod_{i=1}^r n_i!} x_1^{n_1} \cdots x_r^{n_r}.$$

- The number of derangements of a set of size n is

$$n! \left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} \right).$$

Finally, we take this opportunity to state two useful approximations to $n!$:

- Stirling's formula $n! \sim \sqrt{2\pi n} n^n e^{-n}$, which is to say that as n increases $\sqrt{2\pi n} n^n e^{-n} / n! \rightarrow 1$.
- Robbins improved formula* says that

$$\exp\left(-\frac{1}{12n}\right) < \sqrt{2\pi n} n^n e^{-n} / n! < \exp\left(\frac{-1}{12n+1}\right).$$

Checklist of Terms for Chapter 3

- 3.2 permutation
 - multinomial coefficient
 - binomial coefficient
- 3.3 combinations
- 3.4 inclusion–exclusion
 - derangement
- 3.5 recurrence relation
 - Pascal's triangle
- 3.6 generating function
 - exponential generating function
 - convolution
 - binomial theorems
 - coupon collecting
 - multinomial theorem

*Named for H.E. Robbins (1915–2001), famous for the book *What Is Mathematics?*, co-authored with R. Courant.

WORKED EXAMPLES AND EXERCISES

3.9 Example: Railway Trains

Suppose you work in a marshalling yard. Given n similar wagons and r similar locomotives, how many distinct ways are there of making up r trains if empty trains are allowed, and all the wagons are used, and either:

- (a) The wagons and locomotives bear distinct numbers?
or
(b) The wagons are anonymous, but the locomotives are numbered?
or
(c) Wagons and locomotives are anonymous?

What are the corresponding numbers if each train must contain at least m wagons?

Solution (a) First consider how many ways the trains may be sent out of the yard. This is just the number of permutations of $n + r$ objects with the one constraint that the first object has to be a locomotive! The number is $r(n + r - 1)!$. As regards just making up trains, we do not require an ordering of the locomotives, so we divide by $r!$ to get the answer

$$\frac{r(n + r - 1)!}{r!} = \frac{(n + r - 1)!}{(r - 1)!}$$

(b) If the wagons are not numbered, then we cannot distinguish the $n!$ permutations of wagons, so the answer is

$$\frac{(n + r - 1)!}{n!(r - 1)!} = \binom{n + r - 1}{r - 1},$$

which could alternatively have been obtained by observing that we can make up the trains by choosing r places to intersperse the locomotives with the wagons, given that one locomotive must be at the front.

(c) If neither locomotives nor wagons bear numbers, then the number of ways of making up r trains is the number of distinct partitions of n into at most r nonnegative integers, denoted by $p_r(n)$.

For example, the number 5 has altogether seven integral partitions

$$5 = 4 + 1 = 3 + 2 = 3 + 1 + 1 = 2 + 2 + 1 = 2 + 1 + 1 + 1 = 1 + 1 + 1 + 1 + 1$$

so $p_\infty(5) = 7$; of these, four are of three or fewer integers, so $p_3(5) = 4$. There are no simple expressions for $p_r(n)$, but we use the following result, which we state without proof.

Theorem The number of partitions of n with at most k parts is the same as the number of partitions of n in which no part is greater than k .

Now, if we define the generating function

$$g_r(x) = \sum_{n=1}^{\infty} p_r(n)x^n,$$

we can see that

$$1 + g_r(x) = \frac{1}{1-x} \frac{1}{1-x^2} \cdots \frac{1}{1-x^r}$$

by first expanding the right-hand side (by the negative binomial theorem 3.6.12), and then observing that the term x^n arises just as often as n can be written as the sum of positive integers, no one of which is greater than r .

(b') If each train must contain at least m wagons then we require $mr \geq n$. In this case, first suppose the wagons do not have numbers. We attach m to each train in essentially one way, leaving $n - mr$ to be distributed in any way. Using (b), we get the answer $\binom{n-rm+r-1}{r-1}$ ways.

(a') If the wagons have numbers then any of the $n!$ permutations is distinct, giving $n! \binom{n-rm+r-1}{r-1}$ ways.

(c') If neither the locomotives nor wagons have numbers and each train must contain m wagons at least, then we require the number of partitions of m into at most r integers, all of which are not less than m . This is the same as the number of partitions of $n - mr$ into at most r integers, that is, $p_r(n - mr)$, which is the coefficient of x^{n-mr} in $g_r(x)$.

Remark The results about partitions of n are not used elsewhere. They are included as an example of the power of generating functions, which are used extensively throughout probability and related subjects.

- (1) **Exercise** How many ways of making up the r trains are there if you do not have to use all the wagons?
- (2) **Exercise** In how many ways can five oranges be distributed among seven boys? How many ways are there on the more realistic assumption that boys are indistinguishable?
- (3) **Exercise** The hard disk on your PC stores 10 Tb. In how many ways can you divide it up so that no section has less than 3 Tb? (Use only integral multiples of 1 Tb.)

3.10 Example: Genoese Lottery

Suppose that n cards in an urn each bear one of n consecutive integers; all the cards have a different number. Five cards are randomly selected from the urn without replacement. What is the probability that their numbers can be arranged as a run of three consecutive numbers and a nonadjacent run of two consecutive numbers? (For example, 34578 or 23789.)

Solution There are $\binom{n}{5}$ ways of choosing a set of five numbers, and these are assumed to be equally likely. We must count the number of sets providing the two required runs. If the run of three is first and it starts at k , then the run of two may start at $n - k - 4$ places ($1 \leq k \leq n - 5$). Hence, the runs may occur in this order in

$$\sum_{k=1}^{n-5} n - k - 4 = \sum_{j=1}^{n-5} j = \frac{1}{2}(n-5)(n-4)$$

ways.

The run of two is first in another $\frac{1}{2}(n-5)(n-4)$ cases, so the required probability is $(n-5)(n-4)/\binom{n}{5}$.

Remark This lottery is called Genoese because it was introduced by a member of the senate of Genoa named Benedetto Gentile in 1623. It was used to raise money in 1757 by the treasury of Louis XV of France on the advice of Casanova, who had a licence to sell tickets at 6% commission. The treasury also consulted d'Alembert; this was shortly after he had asserted that the chance of HH or TT in two spins of a fair coin is $\frac{2}{3}$.

When Frederick II of Prussia used this lottery to raise money for his treasury, he asked Euler to calculate the odds again.

- (6) **Exercise** As above, n cards bear n consecutive integers. Find the probability that:
- If three cards are drawn, their numbers can form a run of length three.
 - If three cards are drawn, their numbers include a run of length exactly two.
 - If five cards are drawn, their numbers include two runs of length exactly two.

3.11 Example: Ringing Birds

A wood contains n birds, none of which is ringed. Each day one bird is caught, ringed (if it does not already have a ring), and released. Each bird is equally likely to be caught on each day. This procedure is repeated on r successive days, $r \geq n$. Show that the probability of ringing all the birds is $p(r, n)$ where

$$p(r, n) = \sum_{j=0}^n (-1)^j \binom{n}{j} \left(1 - \frac{j}{n}\right)^r$$

Solution I The total number of outcomes (birds being distinguishable), is n^r .

Let $N(r, n)$ be the number of outcomes in which every bird is caught at least once. Because birds are distinct, some bird may be called the first bird. Let $N_k(r, n)$ be the number of outcomes in which the first bird is caught exactly k times, and all the birds are caught at least once. There are $\binom{r}{k}$ ways of choosing the k days on which the first bird is caught, and $N(r-k, n-1)$ ways in which the remaining $n-1$ birds may be caught on the other $r-k$ occasions.

Hence, by Principle 3.1.2,

$$(1) \quad N_k(r, n) = \binom{r}{k} N(r-k, n-1),$$

and, by Principle 3.1.1,

$$(2) \quad N(r, n) = \sum_{k=1}^r N_k(r, n).$$

Of course,

$$(3) \quad N(r, 1) = 1.$$

Substituting (1) into (2) yields a rather complicated recurrence relation. It can be solved by using the exponential generating function:

$$(4) \quad G_n(s) = \sum_{r=1}^{\infty} \frac{s^r}{r!} N(r, n); \quad n > 1.$$

By (3), we have

$$(5) \quad G_1(s) = e^s - 1.$$

Now, multiply (2) by $s^r/r!$ and sum over r [using (1)], to give

$$\begin{aligned} G_n(s) &= \sum_{r=1}^{\infty} \sum_{k=1}^r \frac{s^{r-k} s^k}{(r-k)! k!} N(r-k, n-1) \\ &= \sum_{k=1}^{\infty} \frac{s^k}{k!} G_{n-1}(s) && \text{by the convolution theorem 3.6.6,} \\ &= (e^s - 1)G_{n-1}(s) \\ &= (e^s - 1)^n && \text{on iterating and using (5),} \\ &= e^{ns} \sum_{j=0}^n (-e^{-s})^j \binom{n}{j} && \text{by the binomial theorem} \\ &= \sum_{r=0}^{\infty} \sum_{j=0}^n \frac{(n-j)^r}{r!} (-1)^j \binom{n}{j} s^r && \text{on expanding } e^{(n-j)s}. \end{aligned}$$

Hence,

$$N(r, n) = \sum_{j=0}^n (-1)^j \binom{n}{j} (n-j)^r,$$

using (4) and Theorem 3.6.4. Dividing by n^r yields the required result.

Solution II We may alternatively use the principle of inclusion and exclusion. In the terminology of Section 3.4, “objects” are outcomes and an “object with the k th property” is an outcome in which k birds remain unringed. Now, the number of ways of selecting k birds to remain unringed is $\binom{n}{k}$, and the number of outcomes with k given birds unringed is $(n-k)^r$. Hence, by (3.4.1) and (3.4.3),

$$N(r, n) = n^r - \binom{n}{1} (n-1)^r + \cdots + (-1)^{n-1} \binom{n}{n-1} (n - (n-1))^r,$$

which yields the required result.

- (6) **Exercise** Do the example again using (1.4.8).
 (7) **Exercise** What is the probability that no bird is caught more than twice?
 (8) **Exercise** What is the probability that every bird is caught at least twice?
 (9) **Exercise** As $r \rightarrow \infty$, show that the probability of ringing all the birds converges to one.

3.12 Example: Lottery

In each draw of a lottery, an integer is picked independently at random from the first n integers $1, 2, \dots, n$. What is the probability that in a sample of r successive draws the numbers are drawn in a nondecreasing sequence?

Solution I There are n^r possible sequences in all. We may take any nondecreasing sample sequence of r numbers, together with the first n integers, and arrange these $n + r$ integers as a nondecreasing sequence. Now, place a bar between adjacent different numbers in this sequence, and place a star between adjacent equal numbers. The number to the right of each star is a member of the original sample sequence; there are r stars and $n + r - 1$ places to choose to put them. Furthermore, each such choice corresponds to just one possible sample sequence. By Theorem 3.3.1, there are therefore $\binom{n+r-1}{r}$ nondecreasing sample sequences, so the required probability is $\binom{n+r-1}{r}n^{-r}$.

Solution II Take any sample sequence $s = \{s_1, s_2, \dots, s_r\}$ and add $j - 1$ to s_j to get a unique new sequence $t = \{s_1, s_2 + 1, s_3 + 2, \dots, s_r + r - 1\}$. The sequence t is a selection of r numbers without replacement from $\{1, 2, 3, \dots, n + r - 1\}$, and subtracting $j - 1$ from t_j yields a unique sequence s , which is selected with replacement from $\{1, 2, \dots, n\}$.

Because there are $\binom{n+r-1}{r}$ ways of choosing the sequence t , this is also the number of ways of choosing s .

Remark Observe that this solution includes a proof of Theorem 3.3.4.

- (1) **Exercise** What is the probability that the r numbers are drawn in a strictly increasing sequence?
- (2) **Exercise** Show that the probability that no two drawn numbers are consecutive (i.e., differ by unity) is $r! \binom{n-r+1}{r} n^{-r}$.
- (3) **Exercise** Integers are picked at random from $\{1, 2, \dots, n\}$ until every integer k , $1 \leq k \leq n$, has been drawn at least once. What is the probability that m draws are required to achieve this?

3.13 Example: The Ménages Problem

Suppose that n married couples are seated randomly around a circular table so that men and women alternate.

- (a) Find the number of such seatings in which, for a given set of k couples, the husband and wife are sitting in adjacent seats (some other couples may be adjacent also).
- (b) Hence, deduce the probability that no couple is seated next to each other.

Solution There are two ways of choosing in which seats the women will sit, and $n!$ ways in which they can be arranged in them. The men can be seated in the remaining seats (alternating with the women) in $n!$ ways. Thus, by Principle 3.1.2, there are $2(n!)^2$ possible seating arrangements, which are equally likely by hypothesis.

(a) To count the number of arrangements in which a given set of k couples are adjacent, it is first necessary to count the number of ways of choosing k pairs of adjacent seats for them to occupy. First, consider $2n - k$ chairs in a row. Now, choose k of these [which we may do in $\binom{2n-k}{k}$ ways], and place an extra chair by each of the k chosen chairs. This provides

a choice of k distinct nonoverlapping pairs of adjacent chairs in a row of $2n$. Conversely, for any choice of k disjoint pairs of adjacent chairs from $2n$, we may discard a chair from each pair to give a choice of k from $2n - k$ chairs. This one–one correspondence shows that the number of ways of choosing k disjoint pairs of chairs from $2n$ is just $\binom{2n-k}{k}$.

Now, number the seats round the circular table from 1 to $2n$. By Principle 3.1.1, the number of choices of k disjoint pairs of adjacent seats is the sum of the choices in which $(1, 2n)$ is such a pair [which number $\binom{2n-2-(k-1)}{k-1}$ by the above result], and the choices in which $(1, 2n)$ is not a pair [which number $\binom{2n-k}{k}$].

Hence, the disjoint pairs of seats can be chosen in

$$M_k = \binom{2n-k-1}{k-1} + \binom{2n-k}{k} = \binom{2n}{2n-k} \binom{2n-k}{k}$$

ways.

The k couples to occupy these seats can be chosen in $\binom{n}{k}$ ways, the women's seats chosen in two ways, the k chosen couples arranged in $k!$ ways, and the remaining men and women arranged in $((n-k)!)^2$ ways. Hence, using Principle 3.1.2, the number of seatings for which the k couples are in adjacent seats is

$$(1) \quad S_k = \frac{2n}{2n-k} \cdot \binom{2n-k}{k} \cdot 2k!((n-k)!)^2 = \frac{4n(2n-k-1)!((n-k)!)^2}{(2n-2k)!}$$

(b) Recalling the principle of inclusion and exclusion, we may interpret “an object with the k th property” as “a seating with the k th couple adjacent.” Hence, by using (3.4.3) and (3.4.2), the probability that no couple is adjacent is

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

- (3) **Exercise** Why does it not matter whether we assume the table has a principal seat (head) or not?
- (4) **Exercise** Suppose that n pairs of twins are seated randomly at a round table. What is the probability that no pair of twins sit next to each other? What is the limit of this probability as $n \rightarrow \infty$?
- (5) **Exercise** What is the limit of (2) as $n \rightarrow \infty$?

Remark The problem was first discussed by E. Lucas in 1891 (*ménage* is French for *household*). This method of solution is due to K. Bogart and P. Doyle, *American Mathematical Monthly*, 1986.

3.14 Example: Identity

Show that

$$\sum_{j=0}^r \binom{j+k-1}{k-1} = \binom{r+k}{k}.$$

Solution I Suppose we place r balls randomly in m urns. This is essentially the same as arranging r balls and $m-1$ stars in a row, where the balls between successive stars are placed in successive urns. The positions for the stars may be chosen in $\binom{m+r-1}{r-1}$

ways. Now, in how many of these arrangements do we find b balls in the first urn? The answer is, in just as many ways as the other $r - b$ balls can be put in the other $m - 1$ urns, that is in $\binom{m+r-b-2}{m-2}$ ways. Hence, by Principle 3.1.1,

$$\sum_{b=0}^r \binom{m+r-b-2}{m-2} = \binom{m+r-1}{m-1}.$$

Setting $m - 1 = k$, $r - b = j$ gives the required result (1).

Solution II Multiply each side of (1) by x^r and sum from $r = 0$. By Theorem 3.6.8, we have

$$\frac{1}{1-x} \sum_{r=0}^{\infty} x^r \binom{r+k-1}{k-1} = \sum_{r=0}^{\infty} \binom{r+k}{k} x^r,$$

and by (3.6.10), both sides equal $(1-x)^{-(k+1)}$. The identity (1) follows by Theorem 3.6.4.

Solution III Obviously,

$$\frac{1}{(1-x)^k} = \frac{1}{(1-x)^{k-1}} \frac{1}{1-x}.$$

Equating the coefficient of x^r on each side of this identity yields (1).

Remark Up to a point, these three methods are really the same, but Solution III does illustrate how useful generating functions can be if you happen to hit on an appropriate identity when you need it.

- (2) **Exercise** Show that for any number x and integer j , $\binom{x}{j-1} + \binom{x}{j} = \binom{x+1}{j}$, and hence prove (1) by a fourth method.
- (3) **Exercise** Show that $\sum_{j=0}^n (-1)^j \binom{x}{j} = (-1)^n \binom{x-1}{n}$.
- (4) **Exercise** Show that $\sum_{j=0}^n \binom{x-j}{r} = \binom{x+1}{r+1} - \binom{x-n}{r+1}$.

3.15 Example: Runs

A fair coin is tossed repeatedly. A run of heads is all the heads shown between one tail and the next; if the first toss is a head, there is an opening run up to the first tail; and likewise, there may be a closing run after the last tail. A zero run is no run.

- (a) If the coin is tossed n times, show that the most likely number of runs (of heads, including the opening and closing runs) is $\lfloor \frac{n}{4} \rfloor + 1$ when n is large.
- (b) Also, prove the identity

$$\sum_{i=0}^m \binom{i+k-1}{k-1} \binom{m+k-i}{k} = \binom{2k+m}{m}.$$

Solution (a) There are 2^n possible outcomes. We can choose any outcome with k head runs in the following way. Visualize the n coins in a row. They provide $n + 1$ intervening spaces (including that before the first and that after the last coin). Now, place $2k$ stars

in the spaces, and let the coins between the $(2r + 1)$ th and $(2r + 2)$ th stars be heads ($r = 0, 1, \dots, k - 1$); the other coins are tails. There are $\binom{n+1}{2k}$ ways to do this, so

$$f_k = \mathbf{P}(k \text{ head runs}) = 2^{-n} \binom{n+1}{2k}.$$

Now considering f_k/f_{k+1} , we find

$$\frac{f_k}{f_{k+1}} \geq 1 \quad \text{according as} \quad k \geq \frac{n^2 + n - 2}{4n}.$$

Therefore, the most likely number of runs is the integer next after $\frac{n^2+n-2}{4n}$, which is $\left[\frac{n}{4}\right] + 1$ for large n .

(b) Suppose that the n tosses result in a heads and b tails. First, we divide the a heads into k nonempty groups that form the runs. Imagining the heads in a row, there are $a - 1$ places to put $k - 1$ dividing lines, so the number of ways of getting the k runs of heads is $\binom{a-1}{k-1}$. The head runs alternate with runs of tails, so the b tails are divided into $k + 1$ groups of which the first and last may be empty (providing an opening and/or closing run). If we add an auxiliary tail to each, the same argument as used for head runs shows that the number of ways of arranging the tails is

$$\binom{b+2-1}{k+1-1} = \binom{b+1}{k}.$$

Hence, the number of ways of getting k head runs in n tosses showing a heads is

$$r_k = \binom{a-1}{k-1} \binom{b+1}{k}.$$

The total number of ways of getting k head runs is therefore

$$\begin{aligned} \sum_{a=k}^{n-k+1} r_k &= \sum_{i=0}^{n-2k+1} \binom{i+k-1}{k-1} \binom{n-i-k+1}{k} \\ &= \sum_{i=0}^m \binom{i+k-1}{k-1} \binom{m+k-i}{k}; \quad m = n - 2k + 1, \\ &= \binom{n+1}{2k} \quad \text{by (a)} \\ &= \binom{2k+m}{m} \quad \text{as required.} \end{aligned}$$

Note: In the exercises, n tosses yield a heads and $n - a$ tails.

- (1) **Exercise** What is the probability that the first run is a head run of length k ?
- (2) **Exercise** What is the probability that the last run is a head run of length k ?
- (3) **Exercise** What is the probability that the second run is a head run of length k ?
- (4) **Exercise** What is the probability that the first run is of length k ?

3.16 Example: Fish

A lake contains b bream and c chub. Any fish of each species is equally likely to take a hook.

- (a) If you catch n fish and don't throw any back, what is the probability that you have caught x bream?
- (b) You then return all the fish to the lake (alive) and start fishing again. You now catch m fish. What is the probability that exactly k bream are caught twice?

Solution (a) The question clearly intends us to assume that all possible selections of n fish are equally likely to occur. Then, the number of ways of selecting n fish without repetition is $\binom{b+c}{n}$. The number of ways of catching x bream (and hence also $n - x$ chub) is $\binom{b}{x}\binom{c}{n-x}$, so the required probability is

$$(1) \quad p_x = \frac{\binom{b}{x}\binom{c}{n-x}}{\binom{b+c}{n}}, \quad \max\{0, n-c\} \leq x \leq \min\{b, n\}.$$

(b) We assume that fish do not learn from experience, so that all $\binom{b+c}{m}$ selections of m fish are still equally likely. If x bream were in the first catch, where $k \leq x \leq b$, then the number of ways of selecting m fish, of which k are bream being caught for the second time, is $\binom{x}{k}\binom{b+c-x}{m-k}$. Therefore, the required conditional probability of catching k bream twice, given a first catch of x bream, is

$$p_{k|x} = \binom{x}{k}\binom{b+c-x}{m-k} / \binom{b+c}{m}$$

by the same argument as in (a). Hence, by Theorem 2.1.3, the required unconditional probability is

$$p_k = \sum_x p_{k|x} p_x = \sum_{x=k}^b \frac{\binom{b}{x}\binom{c}{n-x}\binom{x}{k}\binom{b+c-x}{m-k}}{\binom{b+c}{n}\binom{b+c}{m}}$$

- (2) **Exercise** What is the probability of catching x bream if you catch n fish and
 - (a) You throw bream back but not chub?
 - (b) You throw both species back?

Part (a) considers sampling with partial replacement, and part (b) considers sampling with replacement. Discuss the difference in your answers.

- (3) **Exercise** Show that $\sum_{k=0}^{\min\{b,n\}} \binom{b}{k}\binom{c}{n-k} = \binom{b+c}{n}$. [You may want to recall that $(1+x)^b(1+x)^c = (1+x)^{b+c}$.]
- (4) **Exercise** Suppose that as b and c approach ∞ , $b/(b+c) \rightarrow p$ and $c/(b+c) \rightarrow 1-p$. Find the limit of the probabilities in (1) and (2) as b and $c \rightarrow \infty$. Discuss.

3.17 Example: Colouring

Let $K(b, c)$ be the number of different ways in which b indistinguishable balls may be coloured with c different colours. Show that $K(b, c) = K(b - 1, c) + K(b, c - 1)$ and deduce that

$$\sum_{b=0}^{\infty} x^b K(b, c) = (1 - x)^{-c}.$$

Use this to show $K(b, c) = \binom{b+c-1}{c-1}$.

Solution Pick any colour and call it grurple. The number of colourings is the number of ways of colouring the balls which do not colour any grurple, plus the number of ways of colouring which do use grurple. Hence,

$$(1) \quad K(b, c) = K(b - 1, c) + K(b, c - 1).$$

Also, $K(1, c) = c$, and $K(0, c) = 1$ because there are c colours for one ball and only one way of colouring no balls. Now let

$$g_c(x) = \sum_{b=0}^{\infty} x^b K(b, c).$$

Multiply (1) by x^b and sum from $b = 0$ to get

$$(2) \quad g_c(x) = x g_c(x) + g_{c-1}(x).$$

Now using Theorem 3.5.1, we solve (2) to find $g_c(x) = (1 - x)^{-c}$. Furthermore, we may write

$$(1 - x)^{-c} = (1 + x + x^2 + \cdots)(1 + x + x^2 + \cdots) \cdots (1 + x + x^2 + \cdots)$$

Where the right side is the product of c brackets. We get $K(b, c)$ by picking a term from each bracket, and we can say that picking x^k from the i th bracket is like picking k objects of type i . The coefficient $K(b, c)$ of x^b is thus obtained by choosing b objects from c different types of objects with repetition. By Theorem 3.3.4, we have

$$(3) \quad K(b, c) = \binom{c + b - 1}{b}.$$

(4) **Exercise** Let $C(n, k)$ be the number of ways of choosing a set of k objects from n distinct objects. Show that $\sum_{k=0}^n C(n, k)x^k = (1 + x)^n$.

(5) **Exercise** How many nonnegative integer valued solutions for x_1, x_2 , and x_3 does $x_1 + x_2 + x_3 = 20$ have? [For example, $x_1 = 0, x_2 = 4, x_3 = 16$.]

(6) **Exercise** How many positive integer valued solutions does $x_1 + x_2 + x_3 = 20$ have for x_1, x_2 , and x_3 ? [For example, $x_1 = 5, x_2 = 6, x_3 = 9$.]

(7) **Exercise** Show that $K(b, c) = \binom{b+c-1}{c-1}$ by a method different from that in the above solution.

3.18 Example: Matching (Rencontres)

Suppose n different letters are typed with their corresponding envelopes. If the letters are placed at random in the envelopes, show that the probability that exactly r letters match their envelopes is

$$(1) \quad p(n, r) = \frac{1}{r!} \sum_{k=0}^{n-r} \frac{(-1)^k}{k!}.$$

[This problem first surfaced in France during the eighteenth century as a question about coincidences when turning over cards from packs (a kind of French snap).]

Using (3.4.4) gives (1) easily; we display a different method for the sake of variety.

Solution We can suppose that the order of the envelopes is fixed. Let the number of permutations of the letters in which r out of the n letters match their envelopes be $a(n, r)$. Then,

$$p(n, r) = \frac{a(n, r)}{n!}.$$

Suppose we have another letter sealed in its correct envelope. Consider the number A of arrangements that there are of this letter and n letters of which r match their envelopes. We can get this number A in two ways. Either:

- (i) We place the sealed letter in any one of $n + 1$ positions among the n letters to get $(n + 1)a(n, r)$ arrangements:
or:
- (ii) We permute $n + 1$ unsealed letters of which $r + 1$ match and then choose one of the $r + 1$ matching letters to seal, giving $(r + 1)a(n + 1, r + 1)$ arrangements.

The two numbers must both be equal to A , so $(r + 1)a(n + 1, r + 1) = (n + 1)a(n, r)$ and hence, dividing by $(n + 1)!$ we obtain

$$(2) \quad (r + 1)p(n + 1, r + 1) = p(n, r)$$

with $p(n, n) = 1/n!$.

This is a rather interesting recurrence relation, which is solved by standard methods. First iterating (2), we have

$$(3) \quad p(n, r) = \frac{1}{r!} p(n - r, 0).$$

Now define the probability generating function

$$g_n(x) = \sum_{r=0}^n x^r p(n, r); \quad n \geq 1.$$

Multiplying (2) by x^r and summing over r gives

$$g_n(x) = \frac{x^n}{n!} + \sum_{r=1}^n p(r, 0) \frac{x^{n-r}}{(n-r)!}.$$

The sum on the right is a convolution as in Theorem 3.6.6. so multiplying by y^n and summing over n , by Theorem 3.6.6

$$\sum_1^{\infty} y^n g_n(x) = e^{xy} - 1 + e^{xy} \sum_1^{\infty} y^n p(n, 0).$$

Setting $x = 1$ and using Theorem 3.6.10 gives

$$\frac{y}{1-y} = e^y - 1 + e^y \sum_1^{\infty} y^n p(n, 0)$$

so that

$$(4) \quad \sum_1^{\infty} y^n p(n, 0) = \frac{e^{-y}}{1-y} - 1 = \sum_1^{\infty} y^n \sum_{k=0}^n \frac{(-1)^k}{k!}.$$

Hence, by (4) and (3), we get (1).

- (5) **Exercise** Find the probability that exactly $r + s$ matches occur given that at least r matches occur. Show that for large n , it is approximately

$$\frac{1}{(r+s)!} \bigg/ \left(\sum_r^{\infty} \frac{1}{k!} \right).$$

- (6) **Exercise** Show that the probability that the first letter matches its envelope, given that there are exactly r such matches, is $\frac{r}{n}$.
- (7) **Exercise** If a cheque is written for each addressee and these are also placed at random in the envelopes, find:
- The probability that exactly r envelopes contain the correct letter and cheque.
 - The probability that no envelope contains the correct letter and cheque.
 - The probability that every letter contains the wrong letter and the wrong cheque.
- (8) **Exercise** Find the limit of each probability in 7(a), 7(b), and 7(c) as $n \rightarrow \infty$.
- (9) **Exercise** Use 3.4.4 to prove (1) directly.

PROBLEMS

- You have two pairs of red socks, three pairs of mauve socks, and four pairs with a rather attractive rainbow motif. If you pick two socks at random, what is the probability that they match?
- A keen student has a algebra books, b books on boundary layers, and c calculus books. If he places them on one shelf at random, what is the probability that:
 - Books on the same subject are not separated?
 - Books on the same subject are in the usual alphabetical order, but not necessarily adjacent?
 - Books on the same subject are adjacent and in alphabetical order?
- A pack of cards is well shuffled and one hand of 13 cards is dealt to each of four players. Find the probability that:
 - Each player has an ace.
 - At least one player has a complete suit.
 - My hand is void in at least one suit.
 - Some player has all the aces.
 What is the most likely distribution among suits in the dealer's hand?

- 4 **Poker** You are dealt five cards in your hand at poker. What is the probability that you hold:
- (a) One pair? (b) Two pairs?
 (c) A straight? (d) A flush?
 (e) A full house?
- 5 **Birthdays** Assume people are independently equally likely to be born on any day of the year. Given a randomly selected group of r people, of whom it is known that none were born on February 29th, show that the probability that at least two of them have their birthdays either on consecutive days or on the same day is p_r where

$$p_r = 1 - \frac{(365 - r - 1)!}{(365 - 2r)!} 365^{-r+1}, \quad (2r < 365).$$

- Deduce that if $r = 13$, then the probability of at least two such contiguous birthdays is approximately $\frac{1}{2}$, while if $r = 23$ then the probability of at least two such contiguous birthdays is approximately $\frac{9}{10}$.
- 6 You pick an integer at random between zero and 10^5 inclusive. What is the probability that its digits are all different?
- 7 One hundred light bulbs are numbered consecutively from 1 to 100, and are off. They are wired to 100 switches in such a way that the n th switch changes the state (off to on, or on to off) of all the bulbs numbered kn ; $k \geq 1$. If the switches are all thrown successively, how many light bulbs are on? What is the answer if you start with M light bulbs and M switches?
- 8 (a) Show that the product of any r consecutive integers is divisible by $r!$.
 (b) Show that $(k!)!$ is divisible by $(k!)^{(k-1)!}$.
- 9 **Poker Dice** Each die bears the symbols A, K, Q, J, 10, 9. If you roll five such dice, what is the probability that your set of five symbols includes:
- (a) Four aces? (b) Four of a kind? (c) A, K, Q?
- 10 Eight rooks are placed randomly on a chess board (with at most one on each square). What is the probability that:
- (a) They are all in a straight line?
 (b) No two are in the same row or column?
- 11 An urn contains $4n$ balls, n of which are coloured black, n pink, n blue, and n brown. Now, r balls are drawn from the urn without replacement, $r \geq 4$. What is the probability that:
- (a) At least one of the balls is black?
 (b) Exactly two balls are black?
 (c) There is at least one ball of each colour?
- 12 Find the number of distinguishable ways of colouring the faces of a solid regular tetrahedron with:
- (a) At most three colours (red, blue, and green);
 (b) Exactly four colours (red, blue, green, and yellow);
 (c) At most four colours (red, blue, green, and yellow).
- 13 An orienteer runs on the rectangular grid through the grid points (m, n) , $m, n = 0, 1, 2, \dots$ of a Cartesian plane. On reaching (m, n) , the orienteer must next proceed either to $(m + 1, n)$ or $(m, n + 1)$.
- (a) Show the number of different paths from $(0, 0)$ to (n, n) equals the number from $(1, 0)$ to $(n + 1, n)$ and that this equals $\binom{2n}{n}$, where $\binom{k}{r} = \frac{k!}{r!(k-r)!}$.
 (b) Show that the number of different paths from $(1, 0)$ to $(n + 1, n)$ passing through at least one of the grid points (r, r) with $1 \leq r \leq n$ is equal to the total number of different paths from $(0, 1)$ to $(n + 1, n)$ and that this equals $\binom{2n}{n-1}$.
 (c) Suppose that at each grid point the orienteer is equally likely to choose to go to either of the two possible next grid points. Let A_k be the event that the first of the grid points (r, r) , $r \geq 1$,

to be visited is (k, k) . Show that

$$\mathbf{P}(A_k) = \frac{4^{-k}}{2k-1} \binom{2k-1}{k}.$$

- 14** A bag contains b black balls and w white balls. If balls are drawn from the bag without replacement, what is the probability P_k that exactly k black balls are drawn before the first white ball?
By considering $\sum_{k=0}^b P_k$, or otherwise, prove the identity

$$\sum_{k=0}^b \binom{b}{k} / \binom{b+w-1}{k} = \frac{b+w}{w}$$

for positive integers b, w .

- 15** (a) Show that N^{\cdot} , \mathcal{L}^{\cdot} symbols and m^{\cdot} , \cdot symbols may be set out in a line with a^{\cdot} , \cdot at the right-hand end in $\binom{N+m-1}{m-1}$ ways, provided $m \geq 1$.
(b) A rich man decides to divide his fortune, which consists of N one-pound coins, among his m friends. Happily $N > m \geq 1$.
(i) In how many ways can the coins be so divided?
(ii) In how many ways can the coins be so divided if every friend must receive at least one?
(c) Deduce, or prove otherwise, that whenever $N > m \geq 1$,

$$\sum_{k=1}^m \binom{m}{k} \binom{N-1}{k-1} = \binom{N+m-1}{m-1}.$$

- 16** Let N balls be placed independently at random in n boxes, where $n \geq N > 1$, each ball having an equal chance $1/n$ of going into each box. Obtain an expression for the probability P that no box will contain more than one ball. Prove that $N(N-1) < Kn$, where $K = -2 \log P$, and hence that $N < \frac{1}{2} + \sqrt{(Kn + \frac{1}{4})}$.

Now suppose that $P \geq e^{-1}$. Show that $N-1 < 4n/5$ and hence that $Kn < N(N+1)$.

Prove finally that N is the integer nearest to $\sqrt{(Kn + \frac{1}{4})}$ when $P \geq e^{-1}$.

[You may assume that $\log(1-x) < -x$ for $0 < x < 1$, that $\log(1-x) > -x - \frac{3}{2}x^2$ for $0 < x < \frac{4}{5}$, and that $\sum_{r=1}^{N-1} r^2 = N(N-1)(2N-1)/6$.]

- 17** Consider sequences of n integers a_1, a_2, \dots, a_n such that $0 \leq a_i < k$ for each i , where k is a positive integer.
(a) How many such sequences are there?
(b) How many sequences have all a_i distinct?
(c) How many sequences have the property that $a_1 \leq a_2 \leq \dots \leq a_n$?

- 18** Let a_n ($n = 2, 3, \dots$) denote the number of distinct ways the expression $x_1 x_2 \dots x_n$ can be bracketed so that only two quantities are multiplied together at any one time. [For example, when $n = 2$ there is only one way, $(x_1 x_2)$, and when $n = 3$ there are two ways, $(x_1(x_2 x_3))$ and $((x_1 x_2)x_3)$.]

Prove that $a_{n+1} = a_n + a_2 a_{n-1} + a_3 a_{n-2} + \dots + a_{n-2} a_3 + a_{n-1} a_2 + a_n$.

Defining $A(x) = x + a_2 x^2 + a_3 x^3 + \dots$ prove that $(A(x))^2 = A(x) - x$.

Deduce that $A(x) = \frac{1}{2}(1 - (1 - 4x)^{\frac{1}{2}})$, and show that

$$a_n = \frac{1 \cdot 3 \dots (2n-3)}{n!} 2^{n-1}.$$

- 19** **Coupons** Each packet of some harmful and offensive product contains one of a series of r different types of object. Every packet is equally likely to contain one of the r types. If you buy $n \geq r$ packets, show that the probability that you are then the owner of a set of all r types is

$$\sum_{k=0}^r (-1)^k \binom{r}{k} \left(1 - \frac{k}{r}\right)^n.$$

20 Tennis Suppose that $2n$ players enter for two consecutive tennis tournaments. If the draws for each tournament are random, what is the probability that no two players meet in the first round of both tournaments? If n is large, show that this probability is about $e^{-\frac{1}{2}}$.

21 Lotteries Again Suppose that n balls numbered from 1 to n are drawn randomly from an urn. Show that the probability that no two consecutive numbers are actually carried by consecutive balls drawn is

$$1 - \frac{1}{1!} \left(1 - \frac{1}{n}\right) + \frac{1}{2!} \left(1 - \frac{2}{n}\right) - \frac{1}{3!} \left(1 - \frac{3}{n}\right) + \cdots + \frac{(-1)^{n-1}}{n!}.$$

[**Hint:** show that the number of arrangements of $1, 2, \dots, n$ such that at least j pairs of consecutive integers occur is $(n-j)!]$.

22 Runs A fair coin is tossed n times yielding a heads and $n-a$ tails. Show that the probability that there are k head runs and k tail runs (see Example 3.14 for definitions) is $2 \binom{a-1}{k-1} \binom{n-a-1}{k-1} \frac{a!(n-a)!}{n!}$. Deduce that

$$\sum_{k=1}^{a \wedge (n-a)} \binom{a-1}{k-1} \binom{n-a-1}{k-1} / \binom{n-2}{a-1} = 1$$

(where $x \wedge y$ denotes the smaller of x and y).

23 Camelot For obvious reasons Arthur would rather not sit next to Mordred or Lancelot at the Round Table. (There are n seats, and n knights including these three.)

- (a) If the n knights sit at random, what is the probability that Arthur sits next to neither? Does it make any difference whether Arthur sits at random or not?
 (b) If the n knights sit at random on two occasions, what is the probability that no one has the same left-hand neighbour on the two occasions?

24 By considering $(x + x^2 + \cdots + x^n)^r$, show that n indistinguishable objects may be divided into r distinct groups with at least one object in each group in $\binom{n-1}{r-1}$ ways.

25 There are $2n$ balls in an urn; the balls are numbered $1, 2, \dots, 2n$. They are withdrawn at random without replacement. What is the probability that

- (a) For no integer j , the $2j$ th ball drawn bears the number $2j$?
 (b) For no integer j , the ball bearing the number $j+1$ is removed next after the ball bearing the number j ?

Find the limit as $n \rightarrow \infty$ of the probabilities in (a) and (b).

26 A chandelier has seven light bulbs arranged around the circumference of a circle. By the end of a given year, each will have burnt out with probability $\frac{1}{2}$. Assuming that they do so independently, what is the probability that four or more bulbs will have burnt out?

If three bulbs burn out, what is the probability that no two are adjacent?

I decide that I will replace all the dead bulbs at the end of the year only if at least two are adjacent. Find the probability that this will happen. If it does, what is the probability that I will need more than two bulbs?

27 A biased coin is tossed $2n$ times. Show that the probability that the number of heads is the same as the number of tails is $\binom{2n}{n} (pq)^n$. Find the limit of this as $n \rightarrow \infty$.

28 Show that:

- (a) $\sum_0^n (-)^k \binom{n}{k} = 0$; (c) $\sum_{k=0}^{n/2} \binom{n}{2k} = 2^{n-1}$ If n is even;
 (b) $\sum_0^n \binom{n}{k} = 2^n$; (d) $\sum_{k=0}^{n/2} \binom{n}{k} = 2^{n-1}$ If n is even.

29 Observe that

- (i) $(1+x)^m (1+x)^n = (1+x)^{m+n}$
 (ii) $(1-x)^m (1-x)^{-n-2} = (1-x)^{m-n-2}$.

Now show that

- (a) $\sum_{j=0}^k \binom{m}{j} \binom{n}{k-j} = \binom{m+n}{k}$ and (b) $\sum_{k=1}^m (-)^{m-k} \binom{m}{k} \binom{n+k}{n+1} = \binom{n}{m-1}$.

- 30 Show that for $j \leq n/2$, $n^{-n} \sum_{k=0}^j \binom{n}{k} \leq j^{-j} (n-j)^{-(n-j)}$.
- 31 Show that for fixed n , $\binom{n}{k}$ is largest when k is the integer nearest to $\frac{n}{2}$.
- 32 Show that

$$\binom{n}{r} \binom{n+k}{r+2k} \binom{n+2k}{r+k} = \binom{n}{r+k} \binom{n+k}{r} \binom{n+2k}{r+2k},$$

and interpret this in Pascal's triangle.

- 33 Show that $\sum_{k=0}^n \binom{a-k}{b} = \binom{a+1}{b+1} - \binom{a-n}{b+1}$, and deduce that $\sum_{k=0}^n \binom{k+a-1}{a-1} = \binom{n+a}{a}$.
- 34 An urn contains b blue balls and a aquamarine balls. The balls are removed successively at random from the urn without replacement. If $b > a$, show that the probability that at all stages until the urn is empty there are more blue than aquamarine balls in the urn is $(b-a)/(a+b)$.

Why is this result called the ballot theorem?

(Hint: Use conditional probability and induction.)

- 35 The points A_0, A_1, \dots, A_n lie, in that order, on a circle. Let $a_1 = 1$, $a_2 = 1$ and for $n > 2$, let a_n denote the number of dissections of the polygon $A_0A_1 \dots A_n$ into triangles by a set of noncrossing diagonals, A_iA_j .
- (a) Check that $a_3 = 2$ and $a_4 = 5$.
- (b) Show that in each dissection there is a unique i ($1 \leq i \leq n-1$) such that cuts are made along both A_0A_i and A_nA_i .
- (c) Show that $a_n = a_1a_{n-1} + a_2a_{n-2} + \dots + a_{n-2}a_2 + a_{n-1}a_1$.
- (d) If $f(x) = \sum_{n=1}^{\infty} a_n x^n$, show (by considering the coefficient of each power of x) that $(f(x))^2 - f(x) + x = 0$, and show that $f(x) = \frac{1}{2} - \frac{1}{2}\sqrt{1-4x}$.

- 36 Let A, B, C, D be the vertices of a tetrahedron. A beetle is initially at A ; it chooses any of the edges leaving A and walks along it to the next vertex. It continues in this way; at any vertex, it is equally likely to choose to go to any other vertex next. What is the probability that it is at A when it has traversed n edges?
- 37 Suppose that n sets of triplets form a line at random. What is the probability that no three triplets from one set are adjacent?
- 38 Suppose a group of N objects may each have up to r distinct properties b_1, \dots, b_r . With the notation of (3.4.2), show that the number possessing exactly m of these properties is

$$M_m = \sum_{k=0}^{r-m} (-1)^k \binom{m+k}{k} N(b_1, \dots, b_{m+k}).$$

- 39 **The Ménages Problem Revisited** Use the result of Problem 38 to show that the probability that exactly m couples are seated in adjacent seats is

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

- 40 Suppose that N objects are placed in a row. The operation S_k is defined as follows: "Pick one of the first k objects at random and swap it with the object in the k th place." Now perform S_N, S_{N-1}, \dots, S_1 . Show that the final arrangement is equally likely to be any one of the $N!$ permutations of the objects.
- 41 Suppose that n contestants are to be placed in order of merit, and ties are possible. Let $r(n)$ be the number of possible distinct such orderings of the n contestants. (Thus, $r(0) = 0$, $r(1) = 1$, $r(2) = 3$, $r(3) = 13$, and so on.) Show that $r(n)$ has exponential generating function

$$E_r(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} r(n) = \frac{1}{2 - e^x}.$$

[Hint: Remember the multinomial theorem, and consider the coefficient of x^n in $(e^x - 1)^k$.]

- 42 Derangements** (3.4.4 Revisited) Write $\bar{x} = x_1 + x_2 + \cdots + x_n$. Explain why the number of derangements of the first n integers is the coefficient of $x_1 x_2 x_3 \dots x_n$ in

$$\begin{aligned} & (x_2 + x_3 + \cdots + x_n)(x_1 + x_3 + \cdots + x_n) \cdots (x_1 + x_2 + \cdots + x_{n-1}) \\ &= (\bar{x} - x_1)(\bar{x} - x_2) \cdots (\bar{x} - x_n) \\ &= (\bar{x})^n - (\bar{x})^{n-1} \sum x_i + \cdots + (-)^n x_1 x_2 \dots x_n, \end{aligned}$$

and hence deduce the expression for P_n given in (3.4.4).

- 43** (a) Choose n points independently at random on the perimeter of a circle. Show that the probability of there being a semicircular part of that perimeter which includes none of the n points is $n2^{1-n}$.
 (b) Choose n points independently at random on the surface of a sphere. Show that the probability of there being a hemisphere which includes none of the n points is $(n^2 - n + 2)2^{-n}$.
- 44** A large number of students in a lecture room are asked to state on which day of the year they were born. The first student who shares a birthday with someone already questioned wins a prize. Show that, if you were in that audience, your best chance of winning is to be the twentieth person asked.
- 45** The n passengers for an n -seat plane have been told their seat numbers. The first to board chooses a seat at random. The rest, boarding successively, sit correctly unless their allocated seat is occupied, in which case they sit at random. Let p_n be the probability that the last to board finds her seat free. Find p_n , and show that $p_n \rightarrow \frac{1}{2}$, as $n \rightarrow \infty$.

4

Random Variables: Distribution and Expectation

I am giddy, expectation whirls me round.
William Shakespeare, Troilus and Cressida

4.1 Random Variables

In many experiments, outcomes are defined in terms of numbers (e.g., the number of heads in n tosses of a coin) or may be associated with numbers, if we so choose. In either case, we want to assign probabilities directly to these numbers, as well as to the underlying events. This requires the introduction of some new functions.

(1) **Definition** Given a sample space Ω (with \mathcal{F} and $P(\cdot)$), a discrete random variable X is a function such that for each outcome ω in Ω , $X(\omega)$ is one of a countable set D of real numbers. Formally, $X(\cdot)$ is a function with domain Ω and range D , and so for each $\omega \in \Omega$, $X(\omega) = x \in D$, where D is a countable (denumerable) subset of the real numbers. ▲

(2) **Example: Pairs in Poker** How many distinct pairs are there in your poker hand of five cards? Your hand is one outcome ω in the sample space Ω of all possible hands; if you are playing with a full deck, then $|\Omega| = \binom{52}{5}$. The number of pairs X depends on the outcome ω , and obviously $X(\omega) \in \{0, 1, 2\}$, because you can have no more than two pairs. Notice that this holds regardless of how the hand is selected or whether the pack is shuffled. However, this information will be required later to assign probabilities. ●

We always use upper case letters (such as X, Y, T, R , and so on) to denote random variables and lower case letters (x, y, z , etc.) to denote their possible numerical values. You should do the same. Because the possible values of X are countable, we can denote them by $\{x_i; i \in I\}$, where the index set I is a subset of the integers. Very commonly, all the possible values of X are integers, in which case we may denote them simply by x, r, k, j , or any other conventional symbol for integers.

(3) **Definition** (a) If X takes only the values 0 or 1, it is called an indicator or sometimes a Bernoulli trial.

(b) If X takes one of only a finite number of values, then it is called simple. ▲

(4) **Example** Suppose n coins are tossed. Let X_j be the number of heads shown by the j th coin. Then, X_j is obviously zero or one, so we may write $X_j(H) = 1$ and $X_j(T) = 0$.

Let Y be the total number of heads shown by the n coins. Clearly, for each outcome, $\omega \in \Omega$, $Y(\omega) \in \{0, 1, 2, \dots, n\}$. Thus, X_j is an indicator and Y is simple. ●

It is intuitively clear also that $Y = \sum_1^n X_j$.

We discuss the meaning of this and its implications in Chapter 5.

Finally, note that the sample space Ω need not be countable, even though $X(\omega)$ takes one of a countable set of values.

(5) **Example: Darts** You throw one dart at a conventional dartboard. A natural sample space Ω is the set of all possible points of impact. This is of course uncountable because it includes every point of the dartboard, much of the wall, and even parts of the floor or ceiling if you are not especially adroit.

However, your score $X(\omega)$ is one of a finite set of integers lying between 0 and 60, inclusive. ●

4.2 Distributions

Next, we need a function, defined on the possible values x of X , to tell us how likely they are. For each such x , there is an event $A_x \subseteq \Omega$, such that

$$(1) \quad \omega \in A_x \Leftrightarrow X(\omega) = x.$$

Hence, just as the probability that any event A in Ω occurs is given by the probability function $\mathbf{P}(A) \in [0, 1]$, the probability that $X(\omega)$ takes any value x is given by a function $\mathbf{P}(A_x) \in [0, 1]$. (We assume that A_x is in \mathcal{F} .)

For example, if a coin is tossed and X is the number of heads shown, then $X \in \{0, 1\}$ and $A_1 = H$; $A_0 = T$. Hence, $\mathbf{P}(X = 1) = \mathbf{P}(A_1) = \mathbf{P}(H) = \frac{1}{2}$, if the coin is fair.

This function has its own special name and notation. Given Ω , \mathcal{F} , and $\mathbf{P}(\cdot)$:

(2) **Definition** A discrete random variable X has a probability mass function $f_X(x)$ given by

$$f_X(x) = \mathbf{P}(A_x).$$

This is also denoted by $\mathbf{P}(X = x)$, which can be thought of as an obvious shorthand for $\mathbf{P}(\{\omega: X(\omega) = x\})$. It is often called the distribution. ▲

For example, let X be the number of pairs in a poker hand, as discussed in Example 4.1.2. If the hand is randomly selected, then

$$\begin{aligned} f_X(2) &= \mathbf{P}(\{\omega: X(\omega) = 2\}) = |\{\omega: X(\omega) = 2\}| / \binom{52}{5} \\ &= \binom{13}{2} \binom{4}{2} \binom{4}{2} \binom{44}{1} / \binom{52}{5} \simeq 0.048. \end{aligned}$$

Likewise,

$$f_X(1) = \binom{13}{1} \binom{4}{2} \binom{12}{3} \binom{4}{1}^3 / \binom{52}{5} \simeq 0.42,$$

and hence,

$$f_X(0) = 1 - f_X(1) - f_X(2) \simeq 0.53.$$

Returning to Example 4.1.4 gives an example of great theoretical and historical importance.

- (3) **Example 4.1.4 Revisited: Binomial Distribution** The random variable Y takes the value r , if exactly r heads appear in the n tosses. The probability of this event is $\binom{n}{r} p^r q^{n-r}$, where $p = \mathbf{P}(H) = 1 - q$. Hence, Y has probability mass function

$$f_Y(r) = \binom{n}{r} p^r q^{n-r}, \quad 0 \leq r \leq n. \quad \bullet$$

The suffix in $f_X(x)$ or $f_Y(y)$ is included to stress the role of X or Y . Where this is unnecessary or no confusion can arise, we omit it. In the interests of brevity, $f(x)$ is often called simply the *mass function* of X , or even more briefly the p.m.f.

The p.m.f., $f(x) = f_X(x)$, has the following properties: first,

- (4)
$$\begin{aligned} f(x) &\geq 0 \quad \text{for } x \in \{x_i: i \in \mathbb{Z}\} \\ f(x) &= 0 \quad \text{elsewhere.} \end{aligned}$$

That is to say, it is positive for a countable number of values of x and zero elsewhere.

Second, if $X(\omega)$ is finite with probability one, then it is called a *proper* random variable and we have

- (5)
$$\sum_i f(x_i) = \sum_i \mathbf{P}(A_{x_i}) = \mathbf{P}(\Omega) = 1.$$

Third, we have the *Key Rule*

- (6)
$$\mathbf{P}(X \in A) = \sum_{x \in A} f(x).$$

If $\sum_i f(x_i) < 1$, then X is said to be defective or improper. It is occasionally useful to allow X to take values in the extended real line, so that $f_X(\infty)$ has a meaning. In general, it does not.

We remark that any function satisfying (4) and (5) can be regarded as a mass function, in that, given such an $f(\cdot)$, it is quite simple to construct a sample space, probability function, and random variable X , such that $f(x) = \mathbf{P}(X = x)$.

Here are two famous mass functions.

- (7) **Example: Poisson Distribution** Let X be a random variable with mass function

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, 2, \dots\}, \lambda > 0.$$

Then,

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 \quad \text{by Theorem 3.6.9.}$$

Hence, X is proper. This mass function is called the *Poisson distribution* and X is said to be Poisson (or a Poisson random variable), with parameter λ . ●

- (8) **Example: Negative Binomial Distribution** By the negative binomial theorem, for any number q such that $0 < q < 1$, we have

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

Hence, the function $f(r)$ defined by

$$f(r) = \binom{n+r-1}{r} q^r (1-q)^n, \quad r \geq 0,$$

is a probability mass function. Commonly, we let $1 - q = p$. ●

The following function is also useful; see Figure 4.1 for a simple example

Definition A discrete random variable X has a cumulative distribution function $F_X(x)$, where

$$F_X(x) = \sum_{i: x_i \leq x} f(x_i). \quad \blacktriangle$$

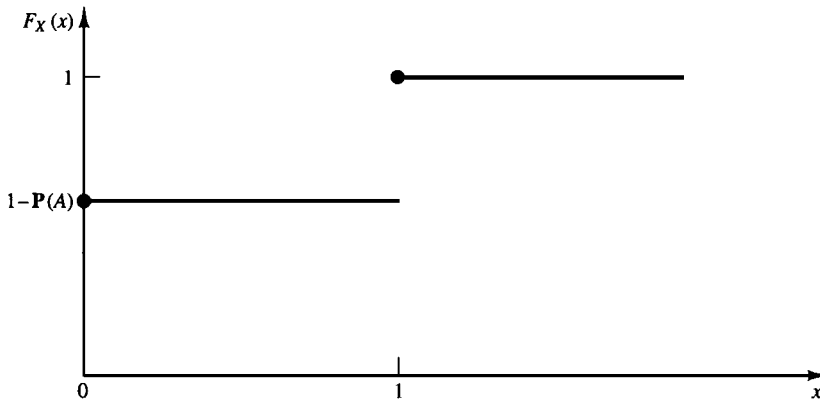


Figure 4.1 The distribution function $F_X(x)$ of the random variable X , which is the indicator of the event A . Thus, the jump at zero is $\mathbf{P}(X = 0) = \mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ and the jump at $x = 1$ is $\mathbf{P}(X = 1) = \mathbf{P}(A)$.

This is also denoted by

$$\mathbf{P}(X \leq x) = \mathbf{P}(\{\omega: X(\omega) \leq x\});$$

it may be referred to simply as the distribution function (or rarely as the c.d.f.), and the suffix X may be omitted. The following properties of $F(x)$ are trivial consequences of the definition (9):

- (10) $F(x) \leq F(y)$ for $x \leq y$.
 (11) $1 - F(x) = \mathbf{P}(X > x)$.
 (12) $\mathbf{P}(a < X \leq b) = F(b) - F(a)$ for $a < b$.

Some further useful properties are not quite so trivial, in that they depend on Theorem 1.5.2. Thus, if we define the event $B_n = \{X \leq x - 1/n\}$, we find that

$$\begin{aligned} \mathbf{P}(X < x) &= \mathbf{P}\left(\bigcup_{n=1}^{\infty} B_n\right) = \mathbf{P}\left(\lim_{n \rightarrow \infty} B_n\right) = \lim_{n \rightarrow \infty} \mathbf{P}(B_n) \quad \text{by Theorem 1.5.2,} \\ &= \lim_{n \rightarrow \infty} F\left(x - \frac{1}{n}\right) = \lim_{y \uparrow x} F(y). \end{aligned}$$

If the random variable X is not defective then, again from (9) (and Theorem 1.5.2), $\lim_{x \rightarrow \infty} F(x) = 1$, and $\lim_{x \rightarrow -\infty} F(x) = 0$.

The c.d.f. is obtained from the p.m.f. by (9). Conversely, the p.m.f. is obtained from the c.d.f. by

$$f(x) = F(x) - \lim_{y \uparrow x} F(y) \quad \text{where } y < x.$$

When X takes only integer values, this relationship has the following simpler more attractive form: for integer x

$$(13) \quad f(x) = F(x) - F(x - 1).$$

- (14) **Example: Lottery** An urn contains n tickets bearing numbers from 1 to n inclusive. Of these, r are withdrawn at random. Let X be the largest number removed if the tickets are replaced in the urn after each drawing, and let Y be the largest number removed if the drawn tickets are not replaced. Find $f_X(x)$, $F_X(x)$, $f_Y(x)$, and $F_Y(x)$. Show that $F_Y(k) < F_X(k)$, for $0 < k < n$.

Solution The number of ways of choosing r numbers less than or equal to x , with repetition allowed, is x^r . Because there are n^r outcomes,

$$F_X(x) = \left(\frac{x}{n}\right)^r, \quad \text{for } 1 \leq x \leq n,$$

when x is an integer. For any real x , $F_X(x) = ([x]/n)^r$, (where $[x]$ denotes the largest integer which is not greater than x). Hence, for integer x and $1 \leq x \leq n$, by (13),

$$f_X(x) = \left(\frac{x}{n}\right)^r - \left(\frac{x-1}{n}\right)^r;$$

and elsewhere $f_X(x)$ is zero.

Without replacement, the number of ways of choosing r different numbers less than or equal to x is $\binom{x}{r}$. Hence, for integer x , and $1 \leq x \leq n$,

$$F_Y(x) = \binom{x}{r} / \binom{n}{r}.$$

Hence, again by (13),

$$f_Y(x) = \left(\binom{x}{r} - \binom{x-1}{r} \right) / \binom{n}{r} = \binom{x-1}{r-1} / \binom{n}{r},$$

which is of course obvious directly. Furthermore,

$$\begin{aligned} F_Y(k) &= \frac{k!(n-r)!}{(k-r)!n!} < \left(\frac{k}{n}\right)^r \quad \text{for } 1 < k < n \\ &= F_X(k). \end{aligned}$$

Because real valued functions of random variables are random variables, they also have probability mass functions.

(15) Theorem *If X and Y are random variables such that $Y = g(X)$, then Y has p.m.f. given by*

$$\sum_{x:g(x)=y} f_X(x).$$

Proof

$$f_Y(y) = \mathbf{P}(g(X) = y) = \sum_{x:g(x)=y} \mathbf{P}(X = x) = \sum_{x:g(x)=y} f_X(x). \quad \blacksquare$$

(16) Example Let X have mass function $f(x)$. Find the mass functions of the following functions of X .

- (a) $-X$
- (b) $X^+ = \max\{0, X\}$
- (c) $X^- = \max\{0, -X\}$
- (d) $|X| = X^+ + X^-$
- (e) $\text{sgn } X = \begin{cases} \frac{X}{|X|}, & X \neq 0, \\ 0, & X = 0. \end{cases}$

Solution Using Theorem 15 repeatedly, we have:

- (a) $f_{-X}(x) = f_X(-x).$
- (b) $f_{X^+}(x) = \begin{cases} f_X(x); & x > 0. \\ \sum_{x \leq 0} f_X(x); & x = 0. \end{cases}$
- (c) $f_{X^-}(x) = \begin{cases} f_X(-x); & x > 0 \\ \sum_{x \geq 0} f_X(x); & x = 0. \end{cases}$

$$(d) \quad f_{|X|}(x) = \begin{cases} f_X(x) + f_X(-x); & x \neq 0 \\ f_X(0); & x = 0. \end{cases}$$

$$(e) \quad f_{\text{sgn}X}(x) = \begin{cases} \sum_{x>0} f_X(x); & x = 1 \\ f_X(0); & x = 0 \\ \sum_{x<0} f_X(x); & x = -1. \end{cases}$$

Finally, we note that any number m such that $\lim_{x \uparrow m} F(x) \leq \frac{1}{2} \leq F(m)$ is called a *median* of F (or a median of X , if X has distribution F)

4.3 Expectation

- (1) **Definition** Let X be a random variable with probability mass function $f(x)$ such that $\sum_x |x|f(x) < \infty$. The expected value of X is then denoted by $\mathbf{E}(X)$ and defined by

$$\mathbf{E}(X) = \sum_x xf(x).$$

This is also known as the expectation, or mean, or average or first moment of X . ▲

Note that $\mathbf{E}(X)$ is the average of the possible values of X weighted by their probabilities. It can thus be seen as a guide to the location of X , and is indeed often called a location parameter. The importance of $\mathbf{E}(X)$ will become progressively more apparent.

- (2) **Example** Suppose that X is an indicator random variable, so that $X(\omega) \in \{0, 1\}$. Define the event $A = \{\omega: X(\omega) = 1\}$. Then X is the indicator of the event A ; we have $f_X(1) = \mathbf{P}(A)$ and $\mathbf{E}(X) = 0 \cdot f_X(0) + 1 \cdot f_X(1) = \mathbf{P}(A)$. ●

- (3) **Example** Let X have mass function

$$f_X(x) = \frac{4}{x(x+1)(x+2)}, \quad x = 1, 2, \dots$$

and let Y have mass function

$$f_Y(x) = \frac{1}{x(x+1)}, \quad x = 1, 2, \dots$$

Show that X does have an expected value and that Y does not have an expected value.

Solution For any $m < \infty$,

$$\sum_{x=1}^m |x|f_X(x) = \sum_{x=1}^m \frac{4}{(x+1)(x+2)} = 4 \sum_{x=1}^m \left(\frac{1}{x+1} - \frac{1}{x+2} \right) = 2 - 4(m+2)^{-1},$$

by successive cancellations in the sum. Hence, the sum converges as $m \rightarrow \infty$, and so because $X > 0$, $\mathbf{E}(X) = 2$. However, for the random variable Y ,

$$\sum_x |x| f_Y(x) = \sum_{x=1}^{\infty} \frac{1}{x+1},$$

which is not finite. ●

Notice that the condition $\sum |x| f(x) < \infty$ amounts to $\mathbf{E}(X^+) + \mathbf{E}(X^-) < \infty$ (use Example 4.2.16 to see this). A little extension of Definition 1 is sometimes useful. Thus, if $\mathbf{E}(X^-) < \infty$ but $\mathbf{E}(X^+)$ diverges, then we may define $\mathbf{E}(X) = +\infty$. With this extension in Example 3, $\mathbf{E}(Y) = \infty$. Likewise, if $\mathbf{E}(X^+) < \infty$ but $\mathbf{E}(X^-)$ diverges, then $\mathbf{E}(X) = -\infty$. If both $\mathbf{E}(X^+)$ and $\mathbf{E}(X^-)$ diverge, then $\mathbf{E}(X)$ is undefined.

In general, real valued functions of random variables are random variables having a mass function given by Theorem 4.2.15. They may therefore have an expected value. In accordance with Example 3, if $Y = g(X)$, then by definition

$$\mathbf{E}(g(X)) = \sum_i y_i f_Y(y_i).$$

We used this with Example 4.2.16(b) in observing that

$$\mathbf{E}(X^+) = \sum_{x>0} x f_X(x).$$

This was easy because it was easy to find the mass function of X^+ in terms of that of X . It is not such an immediately attractive prospect to calculate (for example) $\mathbf{E}(\cos(\theta X))$ by first finding the mass function of $\cos(\theta X)$. The following theorem is therefore extremely useful.

(4) **Theorem** *Let X be a random variable with mass function $f(x)$, and let $g(\cdot)$ be a real valued function defined on \mathbb{R} . Then,*

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

whenever $\sum_x |g(x)| f(x) < \infty$.

Proof Let (g_j) denote the possible values of $g(X)$, and for each j define the set $A_j = \{x: g(x) = g_j\}$. Then $\mathbf{P}(g(X) = g_j) = \mathbf{P}(X \in A_j)$, and therefore, provided all the following summations converge absolutely, we have

$$\begin{aligned} \mathbf{E}(g(X)) &= \sum_j g_j \mathbf{P}(g(X) = g_j) = \sum_j g_j \sum_{x \in A_j} f(x) \\ &= \sum_j \sum_{x \in A_j} g(x) f(x), \quad \text{because } g(x) = g_j \quad \text{for } x \in A_j, \\ &= \sum_x g(x) f(x), \quad \text{because } A_j \cap A_k = \emptyset \quad \text{for } j \neq k. \end{aligned}$$
■

(5) **Example** Let X be Poisson with parameter λ . Find $\mathbf{E}(\cos(\theta X))$.

Solution First, recall de Moivre's Theorem that $e^{i\theta} = \cos \theta + i \sin \theta$, where i is an imaginary square root of -1 . Now, by Theorem 4,

$$\begin{aligned} \mathbf{E}(\cos(\theta X)) &= \sum_{k=0}^{\infty} \cos(k\theta) e^{-\lambda} \lambda^k / k! \\ &= \operatorname{Re} \left(\sum_{k=0}^{\infty} e^{ik\theta} e^{-\lambda} \lambda^k / k! \right), \quad \text{where } \operatorname{Re}(z) \text{ is the real part of } z, \\ &= \operatorname{Re}(\exp(\lambda e^{i\theta} - \lambda)) \\ &= e^{\lambda(\cos \theta - 1)} \cos(\lambda \sin \theta), \quad \text{using de Moivre's Theorem again.} \quad \bullet \end{aligned}$$

Now, we can use Theorem 4 to establish some important properties of $\mathbf{E}(\cdot)$.

(6) **Theorem** Let X be a random variable with finite mean $\mathbf{E}(X)$, and let a and b be constants. Then:

- (i) $\mathbf{E}(aX + b) = a\mathbf{E}(X) + b$;
- (ii) If $\mathbf{P}(X = b) = 1$, then $\mathbf{E}(X) = b$;
- (iii) If $\mathbf{P}(a < X \leq b) = 1$, then $a < \mathbf{E}(X) \leq b$;
- (iv) If $g(X)$ and $h(X)$ have finite mean, then

$$\mathbf{E}(g(X) + h(X)) = \mathbf{E}(g(X)) + \mathbf{E}(h(X)).$$

Proof (i) First, we establish the necessary absolute convergence:

$$\sum_x |ax + b|f(x) \leq \sum_x (|a||x| + |b|)f(x) = |a| \sum_x |x|f(x) + |b| < \infty,$$

as required. Hence, by Theorem 4.

$$\mathbf{E}(aX + b) = \sum_x (ax + b)f(x) = a \sum_x xf(x) + b = a\mathbf{E}(X) + b.$$

(ii) Here, X has mass function $f(b) = 1$, so by definition $\mathbf{E}(X) = bf(b) = b$.

(iii) In this case, $f(x) = 0$ for $x \notin (a, b]$, so

$$\mathbf{E}(X) = \sum_x xf(x) \begin{cases} \leq \sum_x bf(x) = b; \\ > \sum_x af(x) = a. \end{cases}$$

(iv) Because $|g(x) + h(x)| \leq |g(x)| + |h(x)|$, absolute convergence is quickly established. Hence, by Theorem 4,

$$\begin{aligned} \mathbf{E}(g(X) + h(X)) &= \sum_x (g(x) + h(x))f(x) = \sum_x g(x)f(x) + \sum_x h(x)f(x) \\ &= \mathbf{E}(g(X)) + \mathbf{E}(h(X)). \quad \blacksquare \end{aligned}$$

The following simple corollary is of some importance.

(7) **Theorem** If $\mathbf{E}(X)$ exists, then $(\mathbf{E}(X))^2 \leq (\mathbf{E}(|X|))^2 \leq \mathbf{E}(X^2)$.

Proof First, note that $(|X| - \mathbf{E}(|X|))^2 \geq 0$. Hence, by Theorem 6(iii),

$$\begin{aligned} 0 &\leq \mathbf{E}((|X| - \mathbf{E}(|X|))^2) = \mathbf{E}(|X|^2) - (\mathbf{E}(|X|))^2, \quad \text{by Theorem 6(iv) and 6(ii)} \\ &= \mathbf{E}(X^2) - (\mathbf{E}(|X|))^2, \end{aligned}$$

which proves the second inequality. Also, $|X| - X \geq 0$, so by 6(iv) and 6(iii)

$$\mathbf{E}(X) \leq \mathbf{E}(|X|),$$

which proves the first inequality. ■

(8) **Example: Uniform Distribution** Recall that an urn contains n tickets numbered from 1 to n . You take one ticket at random; it bears the number X . Find $\mathbf{E}(X)$ and $\mathbf{E}(X^2)$, and verify that Theorem 7 holds explicitly.

Solution The mass function of X is $\mathbf{P}(X = k) = 1/n$. (Because it distributes probability evenly over the values of X , it is called the *uniform distribution*.) Hence,

$$\begin{aligned} \mathbf{E}(X) &= \sum_{x=1}^n \frac{x}{n} = \frac{1}{n} \sum_{x=1}^n \frac{1}{2}(x(x+1) - x(x-1)) \\ &= \frac{1}{2}(n+1) \quad \text{by successive cancellation.} \end{aligned}$$

Likewise, using Theorems 4 and 6(iv),

$$\begin{aligned} \mathbf{E}(X^2) + \mathbf{E}(X) &= \sum_{x=1}^n \frac{x^2 + x}{n} = \frac{1}{n} \sum_{x=1}^n \frac{1}{3}(x(x+1)(x+2) - (x-1)x(x+1)) \\ &= \frac{1}{3}(n+1)(n+2) \quad \text{by successive cancellation.} \end{aligned}$$

Hence,

$$\mathbf{E}(X^2) = \frac{1}{6}(n+1)(2n+1) \geq \frac{1}{4}(n+1)^2 = (\mathbf{E}(X))^2. \quad \bullet$$

In practice, we are often interested in the expectations of two particularly important collections of functions of X ; namely, $(X^k; k \geq 1)$ and $([X - \mathbf{E}(X)]^k; k \geq 1)$.

(9) **Definition** Let X have mass function $f(x)$ such that $\sum_x |x|^k f(x) < \infty$. Then,

- (a) The k th moment of X is $\mu_k = \mathbf{E}(X^k)$.
- (b) The k th central moment of X is $\sigma_k = \mathbf{E}((X - \mathbf{E}(X))^k)$.
- (c) The k th factorial moment of X is $\mu^{(k)} = \mathbf{E}(X(X-1)\dots(X-k+1))$.

In particular, σ_2 is called the variance of X and is denoted by σ^2 , σ_X^2 , or $\text{var}(X)$. Thus,

$$\text{var}(X) = \mathbf{E}((X - \mathbf{E}(X))^2). \quad \blacktriangle$$

Example: Indicators Let X be the indicator of the event A (recall Example 4.3.2). Because $X^k(\omega) = X(\omega)$ for all k and ω , we have

$$\mu_k = \mathbf{E}(X^k) = \mathbf{P}(A).$$

Also, $\text{var}(X) = \mathbf{P}(A)\mathbf{P}(A^c)$, and

$$\mu^{(k)} = \begin{cases} \mathbf{P}(A); & k = 1 \\ 0; & k > 1. \end{cases} \quad \bullet$$

(10) **Example** Show that if $\mathbf{E}(X^2) < \infty$, and a and b are constants then

$$\text{var}(aX + b) = a^2 \text{var}(X).$$

Solution Using Theorem 6(i) and the definition of variance,

$$\text{var}(aX + b) = \mathbf{E}((a(X - \mathbf{E}(X)) + b - b)^2) = \mathbf{E}(a^2(X - \mathbf{E}(X))^2) = a^2 \text{var}(X). \quad \bullet$$

Sometimes the tail of a distribution, $\mathbf{P}(X > x)$, has a simpler form than the mass function $f(x)$. In these and other circumstances, the following theorems are useful.

(11) **Theorem** If $X \geq 0$ and X takes integer values, then $\mathbf{E}(X) = \sum_{x=1}^{\infty} \mathbf{P}(X \geq x)$.

Proof By definition,

$$\mathbf{E}(X) = \sum_{x=1}^{\infty} xf(x) = \sum_{x=1}^{\infty} f(x) \sum_{r=1}^x 1.$$

Because all terms are nonnegative, we may interchange the order of summation to obtain

$$\sum_{x=1}^{\infty} \sum_{r=x}^{\infty} f(r) = \sum_{x=1}^{\infty} \mathbf{P}(X \geq x). \quad \blacksquare$$

This *tail-sum theorem* has various generalizations; we state one.

(12) **Theorem** If $X \geq 0$ and $k \geq 2$, then

$$\mu^{(k)} = \mathbf{E}(X(X-1) \dots (X-k+1)) = k \sum_{x=k}^{\infty} (x-1) \dots (x-k+1) \mathbf{P}(X \geq x).$$

Proof This is proved in the same way as Theorem 11, by changing the order of summation on the right-hand side. \blacksquare

(13) **Example: Waiting—The Geometric Distribution** A biased coin shows a head with probability p or a tail with probability $q = 1 - p$. How many times do you expect to toss the coin until it first shows a head? Find the various second moments of this waiting time.

Solution Let the required number of tosses until the first head be T . Then because they are independent, $\mathbf{P}(T = x) = q^{x-1}p; x \geq 1$. (T is said to have the *geometric* distribution.)

Hence,

$$\begin{aligned}\mathbf{E}(T) &= \sum_{x=1}^{\infty} xq^{x-1}p = \frac{p}{(1-q)^2}, \quad \text{by (3.6.12) with } n = 2, \\ &= \frac{1}{p}.\end{aligned}$$

Alternatively, we can use Theorem 11 as follows. Using the independence of tosses again gives $\mathbf{P}(T > x) = q^x$, so

$$\mathbf{E}(T) = \sum_{x=0}^{\infty} \mathbf{P}(T > x) = \sum_{x=0}^{\infty} q^x = \frac{1}{p}.$$

For the second factorial moment, by Theorem 12

$$\mu^{(2)} = 2 \sum_{x=2}^{\infty} (x-1)q^{x-1} = \frac{2q}{p^2}, \quad \text{by (3.6.12) again.}$$

Hence, the second moment is

$$\mathbf{E}(T^2) = \mathbf{E}(T(T-1)) + \mathbf{E}(T) = \frac{2q}{p^2} + \frac{1}{p} = \frac{1+q}{p^2}.$$

Finally, the second central moment is

$$\sigma_2 = \mathbf{E}((T - \mathbf{E}(T))^2) = \mathbf{E}(T^2) - (\mathbf{E}(T))^2 = \frac{1+q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}. \quad \bullet$$

(14) Example Let X have mass function

$$f_X(x) = \frac{a}{x^2}; \quad x = 1, 2, 3, \dots$$

and Y have mass function

$$f_Y(y) = \frac{b}{y^2}; \quad y = \pm 1, \pm 2, \dots$$

(a) Find a and b .

(b) What can you say about $\mathbf{E}(X)$ and $\mathbf{E}(Y)$?

Solution (a) Because $f_X(x)$ is a mass function

$$1 = \sum_x f_X(x) = a \sum_{x=1}^{\infty} \frac{1}{x^2} = a \frac{\pi^2}{6}.$$

Hence, $a = 6\pi^{-2}$. Likewise, $b = 3\pi^{-2}$.

(b) We have

$$\mathbf{E}(X) = a \sum_{x=1}^{\infty} x f_X(x) = a \sum_{x=1}^{\infty} \frac{1}{x} = \infty.$$

Because $\mathbf{E}(Y^+)$ and $\mathbf{E}(Y^-)$ both diverge, $\mathbf{E}(Y)$ does not exist. \bullet

- (15) **Example: Coupons** Each packet of an injurious product is equally likely to contain any one of n different types of coupon, independently of every other packet. What is the expected number of packets you must buy to obtain at least one of each type of coupon?

Solution Let A_n^r be the event that the first r coupons you obtain do not include a full set of n coupons. Let C_k^r be the event that you have not obtained one of the k th coupon in the first r . Then,

$$A_n^r = \bigcup_{k=1}^n C_k^r.$$

We may calculate:

$$\begin{aligned} \mathbf{P}(C_1^r) &= \left(\frac{n-1}{n}\right)^r, \\ \mathbf{P}(C_i^r \cap C_j^r) &= \left(\frac{n-2}{n}\right)^r; \quad i \neq j, \end{aligned}$$

and, in general, for any set S_j of j distinct coupons

$$\mathbf{P}\left(\bigcap_{i \in S_j} C_i^r\right) = \left(\frac{n-j}{n}\right)^r.$$

Hence, by (1.4.8),

$$\mathbf{P}(A_n^r) = \mathbf{P}\left(\bigcup_{k=1}^n C_k^r\right) = \sum_{j=1}^n (-1)^{j+1} \binom{n}{j} \left(1 - \frac{j}{n}\right)^r,$$

because for each j there are $\binom{n}{j}$ sets S_j .

Now, let R be the number of packets required to complete a set of n distinct coupons. Because A_n^r occurs if and only if $R > r$, we have $\mathbf{P}(R > r) = \mathbf{P}(A_n^r)$. Hence, by Theorem 11,

$$\begin{aligned} \mathbf{E}(R) &= \sum_{r=0}^{\infty} \mathbf{P}(R > r) = \sum_{r=0}^{\infty} \sum_{j=1}^n (-1)^{j+1} \binom{n}{j} \left(1 - \frac{j}{n}\right)^r \\ &= \sum_{j=1}^n (-1)^{j+1} \binom{n}{j} \frac{n}{j} = nu_n, \quad \text{say.} \end{aligned}$$

Now,

$$\begin{aligned} (16) \quad u_{n+1} - u_n &= \sum_{j=1}^{n+1} (-1)^{j+1} \binom{n+1}{j} \frac{1}{j} - \sum_{j=1}^n (-1)^{j+1} \binom{n}{j} \frac{1}{j} \\ &= \frac{(-1)^{n+2}}{n+1} + \sum_{j=1}^n \frac{(-1)^{j+1}}{j} \left(\binom{n+1}{j} - \binom{n}{j} \right) \\ &= \frac{(-1)^{n+2}}{n+1} + \sum_{j=1}^n \frac{(-1)^{j+1}}{n+1} \binom{n+1}{j} \\ &= \frac{1}{n+1}, \end{aligned}$$

because

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

Hence, iterating (16),

$$u_n = \sum_{j=1}^n \frac{1}{n-j+1},$$

so that

$$\mathbf{E}(R) = \sum_{j=1}^n \frac{n}{n-j+1}.$$

In Chapter 5, we discover a much easier method of obtaining this result. ●

4.4 Conditional Distributions

Let Ω be some sample space, X some random variable defined on Ω , and $\mathbf{P}(\cdot)$ a probability function defined on Ω . Now suppose that we are given that some event $B \subseteq \Omega$ occurs, with $\mathbf{P}(B) > 0$. Just as we argued in Chapter 2 that this gives rise to a conditional probability function, so we now conclude that this gives rise to a conditional distribution of X given B .

In fact, using (4.2.1), we write

$$(1) \quad \mathbf{P}(X(\omega) = x|B) = \mathbf{P}(A_x|B) = \mathbf{P}(A_x \cap B)/\mathbf{P}(B) \geq 0,$$

where as usual $A_x = \{\omega: X(\omega) = x\}$. Furthermore, because $\bigcup_x A_x = \Omega$ and $A_x \cap A_y = \phi$, whenever $x \neq y$ we have

$$(2) \quad \sum_x \mathbf{P}(A_x|B) = \sum_x \mathbf{P}(A_x \cap B)/\mathbf{P}(B) = \mathbf{P}(\Omega \cap B)/\mathbf{P}(B) = 1.$$

Hence, the function $f(x|B)$ defined by

$$(3) \quad f(x|B) = \mathbf{P}(A_x|B) = \mathbf{P}(X = x|B)$$

is a probability mass function, in that $f(x|B) \geq 0$ and $\sum_x f(x|B) = 1$. It is the *conditional mass function* of X given B .

(4) **Example** Let X be uniformly distributed on $\{1, 2, \dots, n\}$, and let B be the event that $a \leq X \leq b$, where $1 \leq a < b \leq n$. Find the mass function of X given B .

Solution Obviously,

$$\mathbf{P}(B) = \sum_{i=a}^b \frac{1}{n} = (b-a+1)/n$$

and

$$\mathbf{P}(\{X = k\} \cap B) = \begin{cases} \frac{1}{n}; & a \leq k \leq b \\ 0; & \text{otherwise.} \end{cases}$$

Hence,

$$f(x|B) = \begin{cases} \frac{1}{b-a+1}; & a \leq x \leq b \\ 0; & \text{otherwise} \end{cases}$$

Thus, given that X lies in B , it is uniformly distributed over B . ●

Because $f(x|B)$ is a probability mass function, it may have an expectation. In line with Definition 4.3.1, we require that $\sum_x |x|f(x|B) < \infty$. If this condition is satisfied, then the *conditional expectation* of X given B is denoted by $\mathbf{E}(X|B)$ and defined by

$$(5) \quad \mathbf{E}(X|B) = \sum_x x f(x|B).$$

Expectation and conditional expectation are related by the following exceptionally important result.

(6) **Theorem** *Let X be a random variable with mean $\mathbf{E}(X)$, and let B be an event such that $\mathbf{P}(B)\mathbf{P}(B^c) > 0$. Then,*

$$\mathbf{E}(X) = \mathbf{E}(X|B)\mathbf{P}(B) + \mathbf{E}(X|B^c)\mathbf{P}(B^c).$$

Proof By conditional probability,

$$\begin{aligned} f(x) &= \mathbf{P}(X = x) = \mathbf{P}(\{X = x\} \cap B) + \mathbf{P}(\{X = x\} \cap B^c) \\ &= f(x|B)\mathbf{P}(B) + f(x|B^c)\mathbf{P}(B^c). \end{aligned}$$

Hence,

$$\mathbf{E}(X) = \sum_x x f(x) = \mathbf{P}(B) \sum_x x f(x|B) + \mathbf{P}(B^c) \sum_x x f(x|B^c),$$

as required. ■

More generally, it is shown in exactly the same way that if $(B_i; i \geq 1)$ is a collection of events such that

- (i) $\bigcup_i B_i = \Omega$,
- (ii) $B_i \cap B_j = \phi$; $i \neq j$, and
- (iii) $\mathbf{P}(B_i) > 0$,

then

$$(7) \quad \mathbf{E}(X) = \sum_i \mathbf{E}(X|B_i)\mathbf{P}(B_i)$$

whenever the summation is absolutely convergent.

Finally, we make the small but useful observation that if $A \subseteq B$, then

$$(8) \quad \mathbf{E}(X|A \cap B) = \mathbf{E}(X|A).$$

- (9) **Example** A coin is tossed repeatedly. As usual, for each toss, $\mathbf{P}(H) = p = 1 - q = 1 - \mathbf{P}(H^c)$. The outcome is a sequence of runs of heads alternating with runs of tails; the first run can be of either heads or tails. Let the length of the n th run be R_n . For all k and j , show that $\mathbf{E}(R_{2k+1}) \geq \mathbf{E}(R_{2j})$ and $\text{var}(R_{2k+1}) \geq \text{var}(R_{2j})$, with equality in each case if and only if $p = q = \frac{1}{2}$.

Solution Let X be the number of heads shown before the first appearance of a tail. We know that

$$(10) \quad \mathbf{P}(X = k) = p^k q, \quad k \geq 0.$$

Let us consider the mass function of X conditional on the first toss. Given that the first toss is H , let X' be the further number of tosses before the first tail. By independence,

$$\mathbf{P}(X' = k) = p^k q = \mathbf{P}(X = k).$$

Hence, conditional on H , we have $X = 1 + X'$, and conditional on H^c , we have $X = 0$. Therefore, by Theorem 6,

$$(11) \quad \mathbf{E}(X) = p\mathbf{E}(X|H) + q\mathbf{E}(X|H^c) = p(1 + \mathbf{E}(X')) + 0 = p + p\mathbf{E}(X) \\ \text{because } \mathbf{E}(X) = \mathbf{E}(X').$$

Thus, $\mathbf{E}(X) = p/q$, which of course we could have obtained directly from (10); we chose to do it this way to display the new technique. Likewise, if Y is the number of tails before the first head, $\mathbf{E}(Y) = q/p$.

Now R_{2k+1} is a run of heads if and only if the first toss is a head. Hence, again using independence,

$$\begin{aligned} \mathbf{E}(R_{2k+1}) &= \mathbf{E}(R_{2k+1}|H)p + \mathbf{E}(R_{2k+1}|H^c)q = \mathbf{E}(1 + X')p + \mathbf{E}(1 + Y')q \\ &= \left(1 + \frac{p}{q}\right)p + \left(1 + \frac{q}{p}\right)q = \frac{p}{q} + \frac{q}{p}. \end{aligned}$$

Likewise, R_{2k} is a run of heads if and only if the first toss yields a tail, so

$$\mathbf{E}(R_{2k}) = \left(1 + \frac{p}{q}\right)q + \left(1 + \frac{q}{p}\right)p = 2 \leq \frac{p}{q} + \frac{q}{p},$$

with equality if and only if $p = \frac{1}{2} = q$. [Because $(p - q)^2 > 0$, for $p \neq q$.] Now, $\text{var}(R_{2k}) = \mathbf{E}(R_{2k}^2) - 4$. Arguing as above, and using conditional probability again, yields

$$\mathbf{E}(R_{2k}^2) = q\mathbf{E}((1 + X')^2) + p\mathbf{E}((1 + Y')^2) = q \left(\frac{1+p}{q^2}\right) + p \left(\frac{1+q}{p^2}\right)$$

and so

$$(12) \quad \text{var}(R_{2k}) = \frac{1 + (p - q)^2 - 2pq}{pq}.$$

Likewise,

$$\mathbf{E}(R_{2k+1}^2) = q \left(\frac{1+q}{p^2} \right) + p \left(\frac{1+p}{q^2} \right)$$

and so

$$(13) \quad \text{var}(R_{2k+1}) = \frac{q}{p^2} + \frac{p}{q^2} - 2.$$

Now

$$\text{var}(R_{2k+1}) - \text{var}(R_{2k}) = \frac{p^4 + q^4 + 2p^2q^2 - pq}{p^2q^2} = \frac{(p^3 - q^3)(p - q)}{p^2q^2} \geq 0$$

with equality if and only if $p = q = \frac{1}{2}$. ●

4.5 Sequences of Distributions

If an experiment is repeated indefinitely, it may give rise to a sequence $(F_n(x); n \geq 1)$ of distributions.

- (1) **Definition** Let $f(x)$ be a probability mass function that is nonzero for $x \in D$, and zero for $x \in \mathbb{R} \setminus D = C$. Let $F(x)$ be the corresponding distribution function

$$F(x) = \sum_{x_i \leq x} f(x_i).$$

A sequence of distribution functions $F_n(x)$ is said to converge to $F(x)$ if, as $n \rightarrow \infty$,

$$F_n(x) \rightarrow F(x) \quad \text{for } x \in C. \quad \blacktriangle$$

One special case is important to us; if D is included in the integers, then $F_n(x)$ converges to $F(x)$ if, for all x , $f_n(x) \rightarrow f(x)$ as $n \rightarrow \infty$.

- (2) **Example: Matching revisited** In Example 3.18, we showed that the probability of exactly r matches in n random assignments of letters is

$$p(n, r) = \frac{1}{r!} \sum_{k=0}^{n-r} \frac{(-)^k}{k!} \rightarrow \frac{e^{-1}}{r!}$$

as $n \rightarrow \infty$. This shows that as $n \rightarrow \infty$ the number of matches has a Poisson distribution (with parameter 1) in the limit. ●

- (3) **Example: Ménages Revisited** In Problem 3.38, we found the probability that exactly m couples were adjacent when seated randomly at a circular table (alternating the

sexes) is

$$p_m = \frac{2}{m!} \sum_{k=0}^{n-m} \frac{(-)^k (n-m-k)!(2n-m-k-1)!}{k!(2n-2m-2k)!n!} \rightarrow \frac{2^m}{m!} \sum_{k=0}^{\infty} \frac{(-)^k 2^k}{k!} = \frac{2^m e^{-2}}{m!}$$

as $n \rightarrow \infty$. Thus, the number of adjacent couples is Poisson with parameter 2 in the limit as $n \rightarrow \infty$. ●

Finally, we note that the appearance of the Poisson distribution in Examples 2 and 3 is significant. This distribution commonly arises in limits of this type, and that is one of the reasons for its major importance.

4.6 Inequalities

Calculating the exact probability that X lies in some set of interest is not always easy. However, simple bounds on these probabilities will often be sufficient for the task in hand. We start with a basic inequality.

- (1) **Theorem: Basic Inequality** *If $h(x)$ is a nonnegative function, then, for $a > 0$*

$$\mathbf{P}(h(X) \geq a) \leq \mathbf{E}(h(X))/a.$$

Proof Define the following function of X :

$$I(h \geq a) = \begin{cases} 1 & \text{whenever } h(X) \geq a \\ 0 & \text{otherwise} \end{cases}$$

Observe that I is an indicator, and so by Example 4.3.2, $\mathbf{E}(I) = \mathbf{P}(h(X) \geq a)$. Now, by its construction I satisfies $h(X) - aI \geq 0$, and so by Theorem 4.3.6 [parts (iii) and (iv)],

$$\mathbf{E}(h(X)) \geq a\mathbf{E}(I) = a\mathbf{P}(h(X) \geq a). \quad \blacksquare$$

The following useful inequalities can all be proved using Theorem 1 or by essentially the same method. You should do some as exercises. For any $a > 0$, we have:

Markov's Inequality

- (2) $\mathbf{P}(|X| \geq a) \leq \mathbf{E}(|X|)/a.$

Chebyshev's Inequality[†]

- (3) $\mathbf{P}(|X| \geq a) \leq \mathbf{E}(X^2)/a^2.$

[†] Some writers use the transliteration "Chebyshev". They then have to remember that the second "e" is pronounced as "o".

One-Sided Chebyshev's Inequality

$$(4) \quad \mathbf{P}(X - \mathbf{E}(X) \geq a) \leq \frac{\text{var}(X)}{a^2 + \text{var}(X)}.$$

Generalized Markov Inequality If $h(x)$ is increasing for $x > 0$, even, and nonnegative then

$$(5) \quad \mathbf{P}(|X| \geq a) \leq \mathbf{E}(h(X))/h(a).$$

If X is nonnegative, then

$$(6) \quad \mathbf{P}(X > a) \leq \frac{\mathbf{E}(X)}{a}.$$

If $c > 0$, then

$$(7) \quad \mathbf{P}(X > a) \leq \frac{\mathbf{E}((X + c)^2)}{(a + c)^2},$$

and

$$(8) \quad \mathbf{P}(X > a) \leq \mathbf{E}(\exp(c(X - a))).$$

Here is one important application.

(9) Example Let X be a random variable such that $\text{var}(X) = 0$. Show that X is constant with probability one.

Solution By (3), for any integer $n \geq 1$,

$$(10) \quad \mathbf{P}\left(|X - \mathbf{E}(X)| > \frac{1}{n}\right) \leq n^2 \text{var}(X) = 0.$$

Hence, defining the events $C_n = \{|X - \mathbf{E}(X)| > 1/n\}$, we have

$$\begin{aligned} \mathbf{P}(X \neq \mathbf{E}(X)) &= \mathbf{P}\left(\bigcup_{n=1}^{\infty} C_n\right) = \mathbf{P}\left(\lim_{n \rightarrow \infty} C_n\right) = \lim_{n \rightarrow \infty} \mathbf{P}(C_n) \quad \text{by Theorem 1.5.2} \\ &= 0. \end{aligned} \quad \bullet$$

An important concept that crops up in many areas of pure and applied mathematics is that of *convexity*. We are interested in the following manifestation of this.

(11) Definition A function $g(x)$ (from \mathbb{R} to \mathbb{R}) is called convex if, for all a , there exists $\lambda(a)$ such that

$$(12) \quad g(x) \geq g(a) + \lambda(a)(x - a), \quad \text{for all } x. \quad \blacktriangle$$

If $g(x)$ is differentiable, then a suitable λ is given by $\lambda(a) = g'(a)$ and (12) takes the form

$$(13) \quad g(x) \geq g(a) + g'(a)(x - a).$$

This says that a convex function lies above all its tangents. If g is not differentiable, then there may be many choices for λ ; draw a picture of $g(x) = |x|$ at $x = 0$ to see this. (There are several other definitions of a convex function, all equivalent to Definition 11.) We are interested in the following property of convex functions.

(14) Theorem: Jensen's Inequality Let X be a random variable with finite mean and $g(x)$ a convex function. Then,

$$(15) \quad \mathbf{E}(g(X)) \geq g(\mathbf{E}(X)).$$

Proof Choosing $a = \mathbf{E}(X)$ in (12), we have

$$g(X) \geq g(\mathbf{E}(X)) + \lambda(X - \mathbf{E}(X)).$$

Taking the expected value of each side gives (15). ■

For example, $g(x) = |x|$ and $g(x) = x^2$ are both convex, so $\mathbf{E}(|X|) \geq |\mathbf{E}(X)|$ and $\mathbf{E}(X^2) \geq (\mathbf{E}(X))^2$. This is Theorem 4.3.7. Here is a less trivial example.

(16) Example Let X be a positive random variable. Show that $\mathbf{E}(\log X) \leq \log \mathbf{E}(X)$.

Solution This follows immediately from Jensen's inequality if we can show that $-\log x$ is convex. Fortunately, this is easy, as follows. By definition, for $x > 0$,

$$\begin{aligned} -\log x &= \int_x^1 y^{-1} dy = \int_a^1 y^{-1} dy + \int_x^a y^{-1} dy, \quad \text{for } a > 0, \\ &= -\log a + \int_x^a y^{-1} dy \geq -\log a + \int_x^a a^{-1} dy = -\log a - \frac{1}{a}(x - a), \end{aligned}$$

and this is (12) with $\lambda(a) = -a^{-1}$. ●

Example 16 has many important applications, of which we see more later. Here is one to begin with.

(17) Example: Arithmetic–Geometric Means Inequality Let $(x_i; 1 \leq i \leq n)$ be any collection of positive numbers and $(p_i; 1 \leq i \leq n)$ any collection of positive numbers such that $\sum_i p_i = 1$. Show that

$$(18) \quad p_1 x_1 + p_2 x_2 + \cdots + p_n x_n \geq x_1^{p_1} x_2^{p_2} \cdots x_n^{p_n}.$$

Solution Let X be the random variable with probability mass function $\mathbf{P}(X = x_i) = p_i; 1 \leq i \leq n$. Then, from (16),

$$\begin{aligned} \log \mathbf{E}(X) &= \log(p_1 x_1 + \cdots + p_n x_n) \geq \mathbf{E}(\log X) \\ &= p_1 \log x_1 + \cdots + p_n \log x_n = \log(x_1^{p_1} x_2^{p_2} \cdots x_n^{p_n}). \end{aligned}$$

The result (18) follows because $\log x$ is an increasing function. ●

- (19) **Example: AM/GM Inequality** In the special case when $p_i = 1/n$, $1 \leq i \leq n$, then (18) takes the form

$$\frac{1}{n} \sum_1^n x_i \geq \left(\prod_{i=1}^n x_i \right)^{1/n} . \quad \bullet$$

4.7 Review and Checklist for Chapter 4

We have seen that many experiments have numerical outcomes; and when they do not, the sample space can be usefully mapped to suitable points on the line. Such real-valued functions defined on the sample space are called random variables. This chapter defined some important types of random variable, and considered several important named examples. We introduced the key concepts of distribution, expectation, functions of random variables, conditional distributions, and sequences of distributions. In Table 4.1 we display the distribution, expectation, and variance for some important specific random variables.

SYNOPSIS OF FORMULAE: In general, every random variable has a distribution function

$$F_X(x) = \mathbf{P}(X \leq x).$$

Discrete random variables, taking one of a countable set of possible values, have a probability mass function

$$f_X(x) = \mathbf{P}(X = x).$$

When X is integer valued, we can write

$$f_X(x) = F(x) - F(x - 1).$$

Table 4.1. *Discrete random variables*

Distribution	$f(x)$	Mean	Variance
Uniform	n^{-1} , $1 \leq x \leq n$	$\frac{1}{2}(n + 1)$	$\frac{1}{12}(n^2 - 1)$
Bernoulli	$p^x(1 - p)^{1-x}$, $x \in \{0, 1\}$	p	$p(1 - p)$
Binomial	$\binom{n}{x} p^x(1 - p)^{n-x}$, $0 \leq x \leq n$	np	$np(1 - p)$
Geometric	$(1 - p)^{x-1} p$, $x \geq 1$	p^{-1}	$(1 - p)p^{-2}$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}$, $x \geq 1$	λ	λ
Negative binomial	$\binom{x-1}{k-1} p^k(1 - p)^{x-k}$, $x \geq k$	kp^{-1}	$k(1 - p)p^{-2}$
Hypergeometric	$\frac{\binom{m}{x} \binom{w}{n-x}}{\binom{m+w}{n}}$, $0 \leq x \leq n$	$\frac{nm}{m+w}$	$\frac{nmw(m+w-n)}{(m+w-1)(m+w)^2}$

In any case, we have the

Key Rule:

$$\mathbf{P}(X \in A) = \sum_{x \in A} f_X(x).$$

Expectation: A discrete random variable has an expected value if $\sum_x |x| f_X(x) < \infty$. In this case,

$$\mathbf{E}X = \sum_x x f_X(x).$$

When X is integer valued and nonnegative,

$$\mathbf{E}X = \sum_{x=0}^{\infty} \mathbf{P}(X > x) = \sum_{x=0}^{\infty} (1 - F(x)).$$

Functions: Suppose that discrete random variables X and Y are such that $Y = g(X)$ for some function $g(\cdot)$. Then,

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x).$$

Also,

$$\mathbf{E}Y = \sum_x g(x) f_X(x)$$

[provided that $\sum_x |g(x)| f_X(x) < \infty$.]

It follows that in any case where each side exists:

$$\mathbf{E}[ag(X) + bh(X)] = a\mathbf{E}g(X) + b\mathbf{E}h(X).$$

Variance: For any random variable X , the variance σ^2 is given by

$$\text{var } X = \mathbf{E}(X - \mathbf{E}X)^2 = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \sigma^2 \geq 0.$$

The number $\sigma \geq 0$ is called the *standard deviation*, and in particular we have, for constants a and b ,

$$\text{var}(aX + b) = a^2 \text{var}X.$$

Moments: The k th moment of X is

$$\mu_k = \mathbf{E}X^k, \quad k \geq 1;$$

usually we write $\mu_1 = \mu$. The k th central moment of X is

$$\sigma_k = \mathbf{E}(X - \mathbf{E}X)^k, \quad k \geq 1;$$

usually we write $\sigma_2 = \sigma^2$.

Conditioning: Any event B in Ω may condition a random variable X , leading to a conditional distribution function

$$F_{X|B}(x|B) = \mathbf{P}(X \leq x|B)$$

and a conditional mass function

$$f_{X|B}(x|B) = \mathbf{P}(X = x|B);$$

Key Rule:

$$\mathbf{P}(X \in A|B) = \sum_{x \in A} f_{X|B}(x|B).$$

This distribution may have an expected value, called

Conditional expectation:

$$\mathbf{E}(X|B) = \sum_x x f_{X|B}(x|B).$$

In particular,

$$\mathbf{E}X = \mathbf{E}(X|B)\mathbf{P}(B) + \mathbf{E}(X|B^c)\mathbf{P}(B^c)$$

and if $(B_i; i \geq 1)$ is a partition of Ω

$$\mathbf{E}X = \sum_i \mathbf{E}(X|B_i)\mathbf{P}(B_i).$$

Similarly, a random variable X may condition an event B , so that

$$\mathbf{P}(B|X = x) = \mathbf{P}(B \cap \{X = x\})/\mathbf{P}(X = x),$$

yielding

$$\mathbf{P}(B) = \sum_x \mathbf{P}(B|X = x)f_X(x).$$

Checklist of Terms for Chapter 4

- 4.1 discrete random variable
 - Bernoulli trial
 - indicator
 - simple random variable
- 4.2 probability distribution
 - probability mass function
 - binomial distribution
 - Poisson distribution
 - negative binomial distribution
 - distribution function
- 4.3 expectation and expected value
 - uniform distribution
 - moments
 - variance
 - tail sum
 - geometric distribution
- 4.4 conditional mass function
 - conditional expectation
- 4.5 sequences of distributions

- 4.6 Markov's inequality
- Chebyshev's inequality
- Jensen's inequality
- Arithmetic-geometric means inequality

WORKED EXAMPLES AND EXERCISES

4.8 Example: Royal Oak Lottery

This eighteenth-century lottery paid winners 28 to 1; the chance of winning at any given bet was 2^{-5} , independently of other bets. Gamesters (as usual) complained that the odds were unfair. It is reported by de Moivre (in *Doctrine of Chances*, 1756) that the Master of the Ball maintained that any particular point of the Ball should come up once in 22 throws; he offered to bet on this (at evens) at any time, and did so when required. The seeming contradiction between the 2^{-5} chance at any bet, with 22 throws for any chance to come up, so perplexed the gamesters that they began to think they had the advantage; so they played on and continued to lose.

Explain why there is no contradiction.

Solution Let \mathbf{P} be a point of the Ball. Let T be the number of trials required to yield \mathbf{P} for the first time. At each trial, \mathbf{P} fails to appear with probability $31/32$, and $T > k$ if and only if the first k trials do not yield \mathbf{P} . Hence, by independence,

$$\mathbf{P}(T > k) = \left(\frac{31}{32}\right)^k.$$

Now,

$$\left(\frac{31}{32}\right)^{22} \simeq 0.49 < 0.5.$$

Hence,

$$\mathbf{P}(T \leq 22) > 0.5.$$

[However, note that $\mathbf{E}(T) = \sum_0^\infty \mathbf{P}(T > k) = (1 - \frac{31}{32})^{-1} = 32$.]

Thus, by betting on the event $T \leq 22$, the Master of the Ball was giving himself a better than evens chance of winning. However, if we let W be the profit to the gambler of a \$1 stake wagered on \mathbf{P} turning up, we have $\mathbf{P}(W = 28) = \frac{1}{32}$ and $\mathbf{P}(W = -1) = \frac{31}{32}$. Hence, $\mathbf{E}(W) = \frac{28}{32} - \frac{31}{32} = -\frac{3}{32}$. A loss.

Thus, in the long run, the gambler will surely lose at a rate of nearly 10% of his stake each play. (See Example 4.18 for a proof of this.)

Remark The Master of the Ball was exploiting the fact that the median of the distribution of T is less than its mean. See Problem 4.51 for bounds on this difference.

Note that T has a *geometric* distribution.

- (1) **Exercise** Give an example of a distribution for which the median is larger than the mean.
- (2) **Exercise** Find: (a) $\text{var}(T)$ and (b) $\mu_T^{(k)}$.
- (3) **Exercise** Which of the following strategies gives the gambler a better chance of winning if she takes up the offer of a bet on \mathbf{P} not occurring in 22 trials:
 (a) Making such a bet immediately?
 (b) Waiting for a run of 22 trials during which \mathbf{P} has not appeared?
 (c) Waiting until \mathbf{P} has appeared in consecutive trials and then betting on its nonappearance in the following 22?
- (4) **Exercise** Calculate $\mathbf{P}(T > j + k | T > j)$. Explain the significance of your answer.

4.9 Example: Misprints

Each printed character in a book is misprinted independently with probability p , or is correct with probability $1 - p$. Let n be the number of characters in the book, and let X be the number of misprinted characters.

- (a) Find $\mathbf{P}(X = r)$.
- (b) Show that $\mathbf{E}(X) = np$.
- (c) Suppose that $\mathbf{E}(X)$ is fixed, and let A be the event that $X \neq 0$. Find $\mathbf{E}(X|A)$, and show that as $n \rightarrow \infty$, $\mathbf{E}(X|A) \rightarrow \mathbf{E}(X)/(1 - \exp[-\mathbf{E}(X)])$.

Solution (a) We provide two solutions.

I Because characters are misprinted independently, the probability that r given characters are misprinted and the remaining $n - r$ are correct is $p^r(1 - p)^{n-r}$. Because there are $\binom{n}{r}$ distinct ways of fixing the positions of the r misprints, it follows that

$$(1) \quad \mathbf{P}(X = r) = \binom{n}{r} p^r (1 - p)^{n-r}.$$

Remark This is the *binomial* distribution, which we met in Example 4.2.3. We sometimes denote it by $B(n, p)$.

II Consider the first character, and let M be the event that it is misprinted. Then,

$$\mathbf{P}(X = r) = \mathbf{P}(X = r|M)\mathbf{P}(M) + \mathbf{P}(X = r|M^c)\mathbf{P}(M^c).$$

We write $\mathbf{P}(X = r) = p(n, r)$ and observe that if M occurs, then $X = r$ if and only if there are $r - 1$ misprints in the remaining $n - 1$ characters. Hence,

$$(2) \quad p(n, r) = p(n - 1, r - 1)p + p(n - 1, r)(1 - p)$$

where $p(n, 0) = (1 - p)^n$ and $p(n, n) = p^n$, $n \geq 0$. Now the substitution $p(n, r) = p^r(1 - p)^{n-r}c(n, r)$ gives $c(n, r) = c(n - 1, r - 1) + c(n - 1, r)$, where $c(n, 0) = c(n, n) = 1$, $n \geq 0$.

We already know that this difference equation has the solution $c(n, r) = \binom{n}{r}$, as required (recall Pascal's triangle).

(b) We consider two solutions.

I Let $m(n)$ be the expected number of misprints in n characters. Then, by Theorem 4.4.6, $m(n) = \mathbf{E}(X|M)\mathbf{P}(M) + \mathbf{E}(X|M^c)\mathbf{P}(M^c) = (1 + m(n - 1))p + m(n - 1)(1 - p)$,

where we have used

$$\begin{aligned} \mathbf{E}(X|M) &= \sum_r r\mathbf{P}(X = r|M) = \sum_r rp(n-1, r-1), \\ & \text{because misprints are independent,} \\ &= \sum_r p(n-1, r-1) + \sum_r (r-1)p(n-1, r-1) = 1 + m(n-1). \end{aligned}$$

Hence,

$$m(n) = m(n-1) + p.$$

Obviously, $m(0) = 0$, so this difference equation has solution

$$(3) \quad m(n) = np.$$

II Using (a)

$$\begin{aligned} m(n) &= \sum_{r=0}^n r \binom{n}{r} p^r (1-p)^{n-r} = \sum_{r=1}^n np \frac{(n-1)!}{(r-1)!(n-r)!} p^{r-1} (1-p)^{n-r} \\ &= np \sum_{r=1}^n \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} = np \quad \text{by the binomial theorem.} \end{aligned}$$

(c) By definition,

$$\begin{aligned} \mathbf{E}(X|A) &= \sum_{r=1}^n \frac{\mathbf{P}(X=r)}{\mathbf{P}(X>0)} = np / (1 - (1-p)^n) \\ &= np / \left(1 - \left(1 - \frac{np}{n}\right)^n\right) = \mathbf{E}(X) / \left(1 - \left(1 - \frac{\mathbf{E}(X)}{n}\right)^n\right) \\ &\rightarrow \mathbf{E}(X) / (1 - \exp(-\mathbf{E}(X))) \end{aligned}$$

as $n \rightarrow \infty$.

- (4) **Exercise** Show that $\text{var}(X) = np(1-p)$ by two different methods.
- (5) **Exercise** Show that as $n \rightarrow \infty$, if $\mathbf{E}(X)$ is fixed $\mathbf{P}(X=0) \rightarrow \exp(-\mathbf{E}(X))$.
- (6) **Exercise** Let X have the binomial distribution $\mathbf{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$.
 - (a) For fixed n and p , for what value of k is $\mathbf{P}(X=k)$ greatest?
 - (b) For fixed k and p , for what value of n is $\mathbf{P}(X=k)$ greatest?
- (7) **Exercise** If X has a binomial distribution with parameters n and p , find:
 - (a) The probability that X is even.
 - (b) $\mathbf{E}(\sin^2(\frac{1}{2}\pi X))$.
 - (c) $\mu_X^{(k)}$.

4.10 Example: Dog Bites: Poisson Distribution

- (a) Let X be a binomial random variable with parameters n and p , such that $np = \lambda$. Show that for fixed k , as $n \rightarrow \infty$, with λ fixed,

$$\mathbf{P}(X = k) \rightarrow \frac{1}{k!} \lambda^k e^{-\lambda}.$$

(b) During 1979–1981, in Bristol, 1103 postmen sustained 215 dog bites. A total of 191 postmen were bitten, of whom 145 were bitten just once. Which should be the postman’s motto: “Once bitten, twice shy” or “Once bitten, twice bitten”?

(8) **Solution** (a) Because X is binomial

$$\mathbf{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{n^k} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Now for fixed k , as $n \rightarrow \infty$ with λ fixed,

$$\frac{n-k+j}{n} \rightarrow 1; \quad 1 \leq j \leq k,$$

$$\left(1 - \frac{\lambda}{n}\right)^k \rightarrow 1, \quad \text{and} \quad \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda}.$$

Hence, as $n \rightarrow \infty$,

$$\mathbf{P}(X = k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

(4) **Remark** This is the *Poisson* distribution, which we met in Example 4.2.7.

(b) Suppose you were a postman, and let X be the number of your bites. If dogs bite any postman at random, then X is a binomial random variable with parameters 215 and $(1103)^{-1}$, because it may be thought of as a series of 215 trials in which a “success” is being bitten with probability $(1103)^{-1}$ at each trial, independently of the rest.

Hence,

$$\mathbf{P}(X = 0) = \left(1 - \frac{1}{1103}\right)^{215}$$

and

$$\mathbf{P}(X = 1) = \frac{215}{1103} \left(1 - \frac{1}{1103}\right)^{214}.$$

You may either compute these directly or recognise from (a) that the number of bites you get is approximately Poisson, with parameter $\lambda = \frac{215}{1103} \simeq 0.195$. So,

$$\begin{aligned} \mathbf{P}(X = 0) &\simeq e^{-\lambda} \simeq 0.82 \\ \mathbf{P}(X = 1) &\simeq \lambda e^{-\lambda} \simeq 0.16 \\ \mathbf{P}(X > 1) &\simeq 1 - e^{-\lambda} - \lambda e^{-\lambda} \simeq 0.02. \end{aligned}$$

However, if we pick a postman at random and let X' be the number of bites he sustained, we find that

$$\begin{aligned} \mathbf{P}(X' = 0) &= \frac{912}{1103} \simeq 0.83 \\ \mathbf{P}(X' = 1) &= \frac{145}{1103} \simeq 0.13 \\ \mathbf{P}(X' > 1) &= \frac{46}{1103} \simeq 0.04. \end{aligned}$$

It seems that “once bitten, twice bitten” should be the postman’s motto.

- (5) **Remark** Our conclusion may be given more substance by investigating the extent to which the observed distribution differs from the expected Poisson distribution. Such techniques are known to statisticians as “goodness-of-fit tests,” and an appropriate procedure here would use the χ^2 test. This may be found in textbooks of elementary statistics; the motto is the same.
- (1) **Exercise** If bites are Poisson with parameter λ , what is the probability that you get more than one bite, given that you get at least one?
- (2) **Exercise** If bites are Poisson, what is your expected total number of bites given that you get at least one?
- (3) **Exercise** Let X_n have a binomial distribution with parameters n and p , such that $np = \lambda$, and let A_n be the event that $X_n \geq 1$. If Y is a Poisson random variable with parameter λ , show that as $n \rightarrow \infty$, $\mathbf{P}(X_n = k | A_n) \rightarrow \mathbf{P}(Y = k | Y \geq 1)$.
- (4) **Exercise** Let X have a Poisson distribution with parameter λ . Show that $\mathbf{E}(X) = \lambda$.
 (a) For fixed λ , what value of k maximizes $\mathbf{P}(X = k)$?
 (b) For fixed k , what value of λ maximizes $\mathbf{P}(X = k)$?
- (5) **Exercise** If X has a Poisson distribution with parameter λ , find:
 (a) $\mathbf{E}(e^X)$ (b) $\mathbf{E}(\cos(\pi X))$ (c) $\text{var}(X)$ (d) $\mu_X^{(k)}$.
- (6) **Exercise** If X has a Poisson distribution with parameter λ , show that:
 (a) The probability that X is even is $e^{-\lambda} \cosh \lambda$.
 (b) $\mathbf{E}(|X - \lambda|) = 2\lambda^\lambda e^{-\lambda} / (\lambda - 1)!$, when λ is an integer greater than zero.

4.11 Example: Guesswork

You are trying to guess the value of a proper integer valued random variable X , with probability mass function $f(x)$ (which you know). If you underestimate by y , it will cost you $\$by$; if you overestimate by y , it will cost you $\$ay$. Your guess is an integer; what guess minimizes your expected loss?

Solution If you guess t , then your expected loss is

$$(1) \quad L(t) = a \sum_{x \leq t} (t - x)f(x) + b \sum_{x > t} (x - t)f(x).$$

Substituting $t + 1$ for t in (1) gives an expression for $L(t + 1)$, and subtracting this from (1) gives

$$(2) \quad L(t) - L(t + 1) = a \sum_{x \leq t} f(x) + b \sum_{x > t} f(x) = -aF(t) + b(1 - F(t)). = D(t)$$

(say).

Now $\lim_{x \rightarrow -\infty} D(x) = b$, $\lim_{x \rightarrow \infty} D(x) = -a$, and both $-F(t)$ and $1 - F(t)$ are non-increasing. Therefore, there is a smallest t such that

$$D(t) = L(t) - L(t + 1) \leq 0,$$

and this is the guess that minimizes your expected loss. Hence, denoting this guess by \hat{t} ,

$$\hat{t} = \min \left\{ t: F(t) \geq \frac{b}{a + b} \right\}, \quad \text{by (2).}$$

- (3) **Exercise** Suppose that if you underestimate X you incur a fixed loss $\mathcal{L}b$, whereas if you overestimate X by y it will cost you $\mathcal{L}ay$. Find an expression that determines the guess that minimizes your expected loss.

Find this best guess when

- (a) $\mathbf{P}(X = x) = pq^{x-1}$; $x \geq 1, p = 1 - q > 0$.
 (b) $\mathbf{P}(X = x) = 1/(x(x+1))$; $x \geq 1$.
 (c) $\mathbf{P}(X = x) = 1/(2n+1)$; $-n \leq x \leq n$.
- (4) **Exercise** What is your best guess if
 (a) $L(t) = \mathbf{E}(|X - t|)$?
 (b) $L(t) = \mathbf{E}((X - t)^2)$?
- (5) **Exercise** Icarus Airways sells $m + n$ tickets for its n -seat aeroplane. Passengers fail to show up with probability p independently. Empty seats cost $\$c$, and a passenger with a ticket who cannot fly is paid $\$b$ for being bumped. What choice of m minimizes the airline's expected losses on booking errors?

What level of compensation b would be sufficient to ensure that it was not worthwhile for the airline to overbook at all (for fixed p)? For fixed b , what value of p would entail no overbooking by the airline?

4.12 Example: Gamblers Ruined Again

Alberich and Brunnhilde have a and b gold coins, respectively. They play a series of independent games in which the loser gives a gold piece to the winner; they stop when one of them has no coins remaining. If Alberich wins each game with probability p , find the expected number of games played before they stop. (Assume $p \neq q = 1 - p$.)

- (3) **Solution** Let X_k be the number of games they will play when Alberich's fortune is k , and let $m_k = \mathbf{E}(X_k)$. Clearly, $m_0 = m_{a+b} = 0$ because in each case one player has no coins. If A is the event that Alberich wins the first game, then for $0 < k < a + b$,

$$\mathbf{E}(X_k|A) = 1 + \mathbf{E}(X_{k+1}) = 1 + m_{k+1}$$

because his fortune is then $k + 1$, and succeeding games are independent of A . Likewise, it follows that

$$\mathbf{E}(X_k|A^c) = 1 + \mathbf{E}(X_{k-1}).$$

Hence, by Theorem 4.4.6,

$$m_k = \mathbf{E}(X_k|A)\mathbf{P}(A) + \mathbf{E}(X_k|A^c)\mathbf{P}(A^c) = 1 + pm_{k+1} + qm_{k-1}.$$

Setting

$$m_k = \frac{k}{q - p} + u_k,$$

gives

$$u_k = pu_{k+1} + qu_{k-1}, \quad \text{for } 0 < k < a + b.$$

In particular $u_0 = 0$, and $u_{a+b} = -(a+b)(q-p)^{-1}$. Proceeding as in Example 2.11, using Theorem 2.3.1, shows that

$$(1) \quad u_k = + \frac{(a+b)}{p-q} \frac{1 - \left(\frac{q}{p}\right)^k}{1 - \left(\frac{q}{p}\right)^{a+b}}.$$

- (2) **Exercise** What is the expected number of games played when $p = \frac{1}{2}$?
- (3) **Exercise** Let B be the event that Brunnhilde wins the entire contest. Find a difference equation satisfied by $\mathbf{E}(X_k|B)$. Solve this in the case when $p = \frac{1}{2}$.
- (4) **Exercise** When the first game is over they redivide the $a+b$ coins as follows. All the coins are tossed, one player gets those showing a head, the other gets all those showing a tail. Now they play a series of games as before. What is the expected number to be played until one or other player again has all the coins? What if $p = \frac{1}{2}$?
- (5) **Exercise** Alberich is blackmailing Fafner, so each time he loses his last gold coin, he immediately demands (and gets) one replacement coin, with which to continue gambling. What now is the expected number of games played? What if $p = \frac{1}{2}$?

4.13 Example: Postmen

A and B are postmen. They start work on day 1. The probability that A sustains a dog bite on day n , given that he has not been bitten on any of the preceding days is $p_A(n)$. The corresponding probability for B is $p_B(n)$. Let X_A and X_B , respectively, be the number of days until each sustains his first bite.

- (a) Find $\mathbf{P}(X_A = n)$ and $\mathbf{P}(X_B = n)$.
- (b) A is wary, so $p_A(n)$ decreases as n increases. If

$$p_A(n) = \frac{1}{n+1}; \quad n \geq 1,$$

find $\mathbf{P}(X_A = n)$ and show that $\mathbf{E}(X_A) = \infty$, while $\mathbf{P}(X_A < \infty) = 1$.

- (c) B is complacent, so $p_B(n)$ increases as n increases. If, for some $\lambda < 0$,

$$p_B(n) = 1 - e^{-\lambda n}$$

find $\mathbf{P}(X_B = n)$ and $\mathbf{E}(X_B)$.

- (2) **Solution** (a) Let H_k be the event that A is bitten on the k th day. Then the event that he is bitten for the first time on the n th day is $\bigcap_{1}^{n-1} H_k^c \cap H_n$. Hence,

$$\begin{aligned} \mathbf{P}(X_A = n) &= \mathbf{P}\left(\bigcap_{1}^{n-1} H_k^c \cap H_n\right) = \mathbf{P}\left(H_n \mid \bigcap_{1}^{n-1} H_k^c\right) \mathbf{P}\left(\bigcap_{1}^{n-1} H_k^c\right) \\ &= p_A(n) \mathbf{P}\left(H_{n-1}^c \mid \bigcap_{1}^{n-2} H_k^c\right) \mathbf{P}\left(\bigcap_{1}^{n-2} H_k^c\right) = p_A(n) \prod_{1}^{n-1} (1 - p_A(k)), \end{aligned}$$

on iterating.

Likewise, $\mathbf{P}(X_B = n) = p_B(n) \prod_1^{n-1} (1 - p_B(k))$.

(b) Employing a similar argument

$$\begin{aligned} \mathbf{P}(X_A > n) &= \mathbf{P}\left(\bigcap_{k=1}^n H_k^c\right) = \prod_{k=1}^n (1 - p_A(k)) \\ &= \prod_{k=1}^n \left(1 - \frac{1}{k+1}\right) = \frac{1}{n+1} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Hence, $\mathbf{P}(X_A < \infty) = 1$. Also,

$$\mathbf{P}(X_A = n) = \mathbf{P}(X_A > n-1) - \mathbf{P}(X_A > n) = \frac{1}{n(n+1)},$$

and finally $\mathbf{E}(X_A) = \sum_1^\infty 1/(n+1)$, which diverges. The expected time until A is first bitten is infinite.

(c) By the same argument,

$$\mathbf{P}(X_B \geq n) = \prod_{j=1}^n (1 - p_B(j)) = \prod_{j=1}^n e^{-\lambda j} = e^{-(\lambda/2)n(n+1)}.$$

Hence, $\mathbf{P}(X_B = n) = (1 - e^{-\lambda n})e^{-(\lambda/2)n(n-1)}$ and

$$\mathbf{E}(X_B) = \sum_1^\infty e^{-(\lambda/2)n(n+1)} < \infty.$$

B expects to be bitten in a finite time.

- (1) **Exercise** In both cases (b) and (c), find the probability that the postman is first bitten on the j th day, given that he is bitten on or before day M .
- (2) **Exercise** If A is less wary, so that $p_A(n) = 2/(n+2)$, show that $\mathbf{E}(X_A)$ is now finite, but $\text{var}(X_A)$ diverges.
- (3) **Exercise** In each case (b) and (c), given that the postman has not been bitten during the first m days, find the expected further time until he is bitten.
- (4) **Exercise** If A is extremely wary and $p_A(n) = 1/(n+1)^2$ show that with probability $\frac{1}{2}$, he is never bitten. What is the median of the distribution of X_A in this case? Find the expectation of X_A , given that X_A is finite.

4.14 Example: Acme Gadgets

This company has developed a new product. The demand for it is unknown, but it is assumed to be a random variable X , which is distributed uniformly on $\{0, 1, \dots, N\}$. The gadgets have to be made in advance; each one sold makes a profit of $\$b$, and each one made and left unsold represents a net loss of $\$c$.

How many should be made, to maximize the expected profit?

Solution Suppose that m items are made. Then the total profit (negative profits are interpreted as losses) is

$$Y_m = \begin{cases} bm; & X \geq m \\ bX - c(m - X); & m > X \end{cases}$$

The expected profit is

$$\begin{aligned} \mathbf{E}(Y_m) &= bm\mathbf{P}(X \geq m) + b \sum_{x=1}^{m-1} \frac{x}{N+1} - c \sum_{x=0}^{m-1} \frac{m-x}{N+1}, \\ &= \frac{m}{N+1} \left(b \left(N + \frac{1}{2} \right) - \frac{1}{2}c - \frac{1}{2}m(b+c) \right). \end{aligned}$$

Now,

$$2(N+1)(\mathbf{E}(Y_{m+1}) - \mathbf{E}(Y_m)) = (2N+1)b - c - (2m+1)(b+c),$$

so that the expected profit is largest when $m = \hat{m}$, where

$$\hat{m} = \max \left\{ 0, \left\lceil \frac{Nb - c}{b + c} \right\rceil \right\}.$$

- (1) **Exercise** Suppose that an unsatisfied customer represents a loss of $\$d$. What now is the choice of m which maximizes expected profit?
- (2) **Exercise** Suppose that the unknown demand X is assumed to have a geometric distribution with parameter p . Find the choice of m that maximizes the expected profit.
- (3) **Exercise** Suppose the unknown demand X is a Poisson random variable with parameter λ . Show that the expected profit if they make m items is

$$\lambda(b+c) - \frac{(b+c)\lambda^{m+1}}{m!} \bigg/ \sum_0^m \frac{\lambda^k}{k!} - mc,$$

and that this is maximized by the value of m that minimizes

$$\frac{(b+c)\lambda^{m+1}}{m!} \bigg/ \sum_0^m \frac{\lambda^k}{k!} + mc.$$

4.15 Example: Roulette and the Martingale

Suppose you are playing roulette; the wheel has a zero. The chance of winning on red is $p < \frac{1}{2}$ and you bet at evens; if you win, you gain an amount equal to your stake.

Your first bet is $\$1$ on red. If it wins, you quit; if it loses, your second bet is $\$2$ on red. If it wins you quit, and so on. Your n th bet is $\$2^{n-1}$ so long as you lose; as soon as you win you quit.

- (a) Show that you are certain to win $\$1$ every time you use this system.
- (b) Find the expected size of your winning bet.

Now suppose the house limit is $\$2^L$, so this must be your last bet if you have not already won.

- (c) What is your expected gain when you stop?
 (d) Would you prefer large or small house limits?

Remark This gambling system is the martingale. Avoid it unless you seek ruin!

Solution Let T be the number of spins of the wheel until the outcome is first red.

(a) Your bet on that spin is $\$2^{T-1}$, and because you win, you gain $\$2^{T-1}$. However, your losses on the previous $T - 1$ spins are $\$ \sum_1^{T-1} 2^{k-1} = \$2^{T-1} - 1$. Because

$$\sum_{k=1}^{\infty} \mathbf{P}(T = k) = \sum_{k=1}^{\infty} p(1-p)^{k-1} = 1,$$

this means you are certain to win \$1.

(b) Because your winning bet is $\$2^{T-1}$, it has expected value

$$\mathbf{E}(2^{T-1}) = \sum_{k=1}^{\infty} 2^{k-1} p(1-p)^{k-1} = \infty, \quad \text{since } 2(1-p) > 1.$$

(c) You win \$1 if $1 \leq T \leq L + 1$; otherwise, you lose $\$ \sum_{k=0}^L 2^k$. Hence, your expected gains are $\$ \gamma$, where

$$\gamma = \mathbf{P}(T \leq L + 1) - (2^{L+1} - 1)\mathbf{P}(T > L + 1) = 1 - (2(1-p))^{L+1}.$$

(d) Because your expected losses increase exponentially fast with L , you must hope the casino is sufficiently generous to have low limits.

- (1) **Exercise** What difference does it make to these results if the wheel is fair? (That is, $p = \frac{1}{2}$.)
 (2) **Exercise** With house limit $\$2^L$, what is the expected size of your winning bet, given that you do indeed win? What happens as $L \rightarrow \infty$? (Remember to consider all three cases, $p > \frac{1}{2}$, $p = \frac{1}{2}$, $p < \frac{1}{2}$.)

4.16 Example: Searching

(a) Let X be a positive integer valued random variable such that $f(n) = \mathbf{P}(X = n)$ is nonincreasing as n increases. Suppose that $(g(x); x = 1, 2, \dots)$ is a function, taking positive integer values, such that for any k , $g(x) = k$ for at most one positive integer $x = r_k$.

Show that $\mathbf{E}(g(X)) \geq \mathbf{E}(X)$.

(b) You have lost a key. There are n places in which you might have mislaid it with respective probabilities $(p_k; 1 \leq k \leq n)$. If you search the k th place once, you find the key with probability d_k , if it is indeed there. (You can search any place any number of times.) How do you arrange your searching to minimize the expected time until you find the key? (Searches are successful independently of each other.)

Solution (a) Consider the distribution of $g(X)$. Because $g(X) = k$ for at most one value r_k of X ,

$$\mathbf{P}(g(X) \leq n) = \sum_k \mathbf{P}(g(X) = k) = \sum_k \mathbf{P}(X = r_k) = \sum_k f(r_k)$$

where the final sum contains $m \leq n$ nonzero terms. If these are arranged in decreasing order as $f(r_{k_1}) \geq f(r_{k_2}) \geq \dots \geq f(r_{k_m})$, then

$$\begin{aligned} f(r_{k_1}) &\leq f(1) \\ f(r_{k_2}) &\leq f(2), \end{aligned}$$

and so on. Hence, summing these inequalities yields

$$\mathbf{P}(g(X) \leq n) \leq \mathbf{P}(X \leq n),$$

and so

$$\begin{aligned} \mathbf{E}(g(X)) &= \sum_0^\infty \mathbf{P}(g(X) > n) = \sum_0^\infty (1 - \mathbf{P}(g(X) \leq n)) \\ &\geq \sum_0^\infty (1 - \mathbf{P}(X \leq n)) = \mathbf{E}(X). \end{aligned}$$

(b) The probability that you find the key on the s th search of the r th room is

$$m_{rs} = (1 - d_r)^{s-1} d_r p_r.$$

To see this, note that the key has to be there (with probability p_r) and you have to fail to find it $s - 1$ times before you succeed. Let p_k be the k th largest of the numbers $(m_{rs}; r \geq 1, s \geq 1)$. Then p_k is a probability mass function and $(p_k; k \geq 1)$ is nonincreasing. Take this ordering as an order of search; that is, if $m_{rs} \geq m_{uv}$, then the s th search of the r th place precedes the v th search of the u th place. This search is consistent [the m th search of a given place precedes the $(m + 1)$ st for every m], and $\sum k p_k$ is the expected number of searches required to find the key.

By part (a), any other order yields greater expected duration of the search time, because the function $g(x)$ is a permutation, and thus one–one.

- (1) **Exercise** Show that you can arrange your searches so that the expected time to find the key is finite.
- (2) **Exercise** The key is upstairs with probability $\frac{2}{3}$ or downstairs, with probability $\frac{1}{3}$. Any search upstairs is successful with probability $\frac{1}{4}$ if the key is there; any search downstairs is successful with probability $\frac{3}{4}$ if the key is there. How do you arrange your searches to minimize the expected number of searches?
- (3) **Exercise** Suppose the s th search of the r th room (conditional on $s - 1$ previous unsuccessful searches of this room) discovers the key with probability d_{rs} . How do you order your searches to minimize the expected number of searches?

4.17 Example: Duelling

Pascal and Brianchon fight a series of independent bouts. At each bout, either Pascal is awarded a hit with probability p , or Brianchon is awarded a hit with probability $q = 1 - p$. The first to be awarded two consecutive hits is declared the winner and the duel stops. Let X be the number of bouts fought. Find the distribution and expected value of X . For what value of p is $\mathbf{E}(X)$ greatest?

Solution Let B be the event that Brianchon wins. Then,

$$f_X(n) = \mathbf{P}(\{X = n\} \cap B) + \mathbf{P}(\{X = n\} \cap B^c).$$

For B to occur at the n th bout, he must win the n th and $(n - 1)$ th bouts (with probability q^2), and the preceding $n - 2$ bouts must be awarded alternately to each contestant. The probability of this is $p^{(n/2)-1}q^{(n/2)-1}$ if n is even, or $p^{(n/2)-(1/2)}q^{(n/2)-(3/2)}$ if n is odd, because bouts are independent.

A similar argument applies if B^c occurs, yielding

$$f_X(n) \begin{cases} p^{(n/2)-1}q^{(n/2)-1}(q^2 + p^2) & \text{if } n \text{ is even} \\ p^{(n/2)-(1/2)}q^{(n/2)-(1/2)}(q + p) & \text{if } n \text{ is odd.} \end{cases}$$

The expected value of X is then, by definition,

$$\mathbf{E}(X) = \sum_{j=1}^{\infty} p^{j-1}q^{j-1}(q^2 + p^2)2j + \sum_{j=1}^{\infty} q^j p^j (q + p)(2j + 1).$$

Summing this series is elementary and boring. To get a solution in closed form, it is more fun to argue as follows. Let A_k be the event that Pascal is awarded the k th bout. Then,

$$(1) \quad \mathbf{E}(X) = \mathbf{E}(X|A_1)p + \mathbf{E}(X|A_1^c)q,$$

by conditioning on the outcome of the first bout. Now if Pascal is awarded the first bout but not the second, the state of the duel in respect of the final outcome is exactly the same as if he had lost the first bout, except of course that one bout extra has been fought. Formally, this says

$$\mathbf{E}(X|A_1 \cap A_2^c) = 1 + \mathbf{E}(X|A_1^c).$$

Hence,

$$(2) \quad \mathbf{E}(X|A_1) = \mathbf{E}(X|A_1 \cap A_2)p + \mathbf{E}(X|A_1 \cap A_2^c)q = 2p + q(1 + \mathbf{E}(X|A_1^c)).$$

Likewise,

$$(3) \quad \mathbf{E}(X|A_1^c) = 2q + p(1 + \mathbf{E}(X|A_1)).$$

Solving (2) and (3), and substituting into (1), yields

$$\mathbf{E}(X) = \frac{2 + qp}{1 - qp}.$$

Because $qp = \frac{1}{4} - (p - \frac{1}{2})^2$, this is greatest when $p = \frac{1}{2}$ and then $\mathbf{E}(X) = 3$.

- (4) **Exercise** What is $\mathbf{P}(B)$?
- (5) **Exercise** What is $\mathbf{P}(B^c)$?
- (6) **Exercise** Given that Pascal wins, find the distribution and expected value of the number of bouts.
- (7) **Exercise** Find $\mathbf{P}(A_1|B)$ and $\mathbf{P}(A_2|B)$.
- (8) **Exercise** Find the median number of bouts when $p = \frac{1}{2}$.
- (9) **Exercise** Find $\mathbf{P}(B)$ and the expectation of the number of bouts fought if the winner is required to win *three* consecutive bouts.

- (10) **Exercise** Brianchon suggests that they adopt a different rule for deciding the winner, *viz.*: when first a player has been awarded a total number of bouts two greater than the number of bouts awarded to his opponent, then the match stops and the leading player wins. If $p > q$, do you think Brianchon was wise to suggest this? (Assume he wants to win.) What is the expected duration of this game when $p = q$?

4.18 Binomial Distribution: The Long Run

Let X have a binomial distribution with parameters n and p , where $p = 1 - q$. Show that for $\lambda > 0$ and $\epsilon > 0$,

$$\mathbf{P}(X - np > n\epsilon) \leq \mathbf{E}(\exp[\lambda(X - np - n\epsilon)]).$$

Deduce that as $n \rightarrow \infty$,

$$\mathbf{P}(|X - np| \leq n\epsilon) \rightarrow 1.$$

(You may assume without proof that for any x , $0 < e^x \leq x + e^{x^2}$.)

Solution For $k > np + n\epsilon$, when $\lambda > 0$, we have $\exp(\lambda(k - np - n\epsilon)) > 1$. Hence

$$\begin{aligned} \sum_{k > n(p+\epsilon)} \mathbf{P}(X = k) &< \sum_{k > n(p+\epsilon)} \exp(\lambda(k - np - n\epsilon))\mathbf{P}(X = k) \\ &< \sum_k \exp(\lambda(k - np - n\epsilon))\mathbf{P}(X = k), \quad \text{because } e^x > 0, \\ &= \mathbf{E}(\exp(\lambda(X - np - n\epsilon))). \end{aligned}$$

Now, the left side is just $\mathbf{P}(X > np + n\epsilon)$ and

$$\mathbf{E}(e^{\lambda X}) = \sum_0^n \binom{n}{k} (pe^\lambda)^k q^{n-k} = (q + pe^\lambda)^n,$$

so the right side is

$$\begin{aligned} (pe^{\lambda q} + qe^{-\lambda p})^n e^{-\lambda n\epsilon} &\leq (pe^{\lambda^2 q^2} + qe^{\lambda^2 p^2})^n e^{-\lambda n\epsilon}, \quad \text{because } e^x \leq x + e^{x^2} \\ &\leq \exp(n\lambda^2 - \lambda n\epsilon). \end{aligned}$$

Now, choosing $\lambda = \epsilon/2$ gives

$$\mathbf{P}(X - np > n\epsilon) \leq \exp(-n\epsilon^2/4).$$

Likewise, $\mathbf{P}(X - np < -n\epsilon) \leq \exp(-n\epsilon^2/4)$, so

$$(1) \quad \mathbf{P}(|X - np| > n\epsilon) \leq 2 \exp(-n\epsilon^2/4) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

as required.

- (2) **Exercise** You want to ask each of a large number n of people a question to which the answer “yes” is so embarrassing that many individuals would falsely answer “no”. The answer “no” is not embarrassing. The following procedure is proposed 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 true answer would have been “no” and the coin shows heads, then the answer “yes” is given. Otherwise, people

should respond truthfully. If the number responding “yes” is now Y_n and “yes” is the true answer for a proportion p of the whole population, show that for $\epsilon > 0$

$$\mathbf{P}\left(\left|\frac{Y_n}{n} - \frac{1}{2}(1+p)\right| > \epsilon\right) \leq 2 \exp(-n\epsilon^2/4).$$

Explain the advantages of this procedure.

- (3) **Exercise** Suppose a coin shows a head with probability p , and let X_n be the number of heads in n tosses, and $A_n(\epsilon)$ the event that $|X_n/n - p| > \epsilon$, where $2 > \epsilon > 0$. Show that as $n \rightarrow \infty$,

$$\mathbf{P}\left(\bigcup_n A_n(\epsilon)\right) \rightarrow 0.$$

- (4) **Exercise** Suppose a gambler wins \$28 with probability $\frac{1}{32}$, or loses his stake of \$1 with probability $\frac{31}{32}$ at each trial. Let W_n be his fortune after n such independent trials, and $A_n(\epsilon)$ the event that $|W_n/n + 3/32| > \epsilon$. Show that as $n \rightarrow \infty$, $\mathbf{P}(\bigcup_{n=m}^{\infty} A_n(\epsilon)) \rightarrow 0$. Deduce that his fortune is equal to its initial value on only finitely many occasions, with probability one. (*Hint*: recall Problem 1.24.)

Note: In the following exercise, X is a binomially distributed random variable with parameters n and p .

- (5) **Exercise** Show that for any fixed finite a and b , as $n \rightarrow \infty$, $\mathbf{P}(a < X \leq b) \rightarrow 0$.
 (6) **Exercise** Show that for $a > 0$,

$$\mathbf{P}\left(\left|\frac{X}{n} - p\right| > a\right) \leq \frac{(p(1-p))^{1/2}}{a^2 n} \min\{(p(1-p))^{1/2}, an^{1/2}\}.$$

- (7) **Exercise** (a) Show that if $p = (m-1)/n$ where m is an integer, then

$$\mathbf{E}\left(\left|\frac{X}{n} - p\right|\right) = 2 \binom{n-1}{m-1} p^m (1-p)^{n-m+1}.$$

(b) Find $\text{var}(|X/n - p|)$.

- (8) **Exercise** If $n = 2m$ and $p = \frac{1}{2}$, show that

$$\mathbf{P}(X - m = k) = \binom{2m}{m} \frac{1}{4^m} a(m, k)$$

where, as $m \rightarrow \infty$, $(a(m, k))^m \rightarrow e^{-k^2}$.

Also, show that

$$\frac{1}{2m^{1/2}} < \binom{2m}{m} \frac{1}{4^m} < \frac{1}{(2m+1)^{1/2}}.$$

[You may assume that $|\log(1+x) - x| < x^2$ for small enough x .]

4.19 Example: Uncertainty and Entropy

Let X and Y be simple random variables taking values in the same set $\{x_1, \dots, x_n\}$, with respective probability mass functions $f_X(\cdot)$ and $f_Y(\cdot)$. Show that

- (1) $-\mathbf{E}(\log f_Y(X)) \geq -\mathbf{E}(\log f_X(X)),$

and

$$(2) \quad -\mathbf{E}(\log f_X(X)) \leq \log n,$$

with equality in (1) if and only if $f_Y(\cdot) \equiv f_X(\cdot)$, and equality in (2) if and only if $f_X(x_i) = n^{-1}$ for all x_i . (*Hint*: Show first that $\log x \leq x - 1$.)

Solution By definition, for $x > 0$,

$$(3) \quad -\log x = \int_x^1 y^{-1} dy \geq \int_x^1 dy = 1 - x,$$

with equality if and only if $x = 1$. Hence,

$$\begin{aligned} \mathbf{E}(\log f_X(X)) - \mathbf{E}(\log f_Y(X)) &= \sum_i f_X(x_i) \log f_X(x_i) - \sum_i f_X(x_i) \log f_Y(x_i) \\ &= -\sum_i f_X(x_i) \log[f_Y(x_i)/f_X(x_i)] \\ &\geq -\sum_i f_X(x_i)[1 - f_Y(x_i)/f_X(x_i)] \quad \text{by (3)} \\ &= 0, \end{aligned}$$

with equality iff $f_X(x_i) = f_Y(x_i)$ for all x_i , which proves (1). In particular, setting $f_Y(x_i) = n^{-1}$ yields (2).

Remark It is conventional to denote $-\mathbf{E}(\log f_X(X))$ by $H(X)$ [or alternatively $h(X)$] and the logarithms are taken to base 2. The number $H(X)$ is known as the *uncertainty* or *entropy* of X , and is an essential tool in information theory and communication theory. The result (1) is sometimes called the Gibbs inequality.

(4) **Exercise** Let $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$; $0 \leq x \leq n$. Show that

$$H(X) \leq -n(p \log p + (1-p) \log(1-p)),$$

with equality if $n = 1$.

(5) **Exercise** Let $f_X(x) = pq^{x-1}/(1-q^M)$, for $1 \leq x \leq M$, where $p = 1 - q$. Show that

$$\lim_{M \rightarrow \infty} H(X) = -p^{-1}[p \log p + (1-p) \log(1-p)].$$

(6) **Exercise** Let $Y = g(X)$ be a function of the random variable X . Show that for any $c > 0$

$$H(Y) \leq H(X) \leq c\mathbf{E}(Y) + \log \left(\sum_i e^{-cg(x_i)} \right).$$

When does equality hold?

PROBLEMS

- 1 A box contains 12 sound grapefruit and four that are rotten. You pick three at random.
- Describe the sample space.
 - Let X be the number of sound grapefruit you pick. Find $f_X(x)$ and $\mathbf{E}(X)$.

- (c) A careless miscreant leaves a clear fingerprint of type t . It is known that the probability that any randomly selected person has this type of fingerprint is 10^{-6} . The city has 10^7 inhabitants and a citizen is produced who has fingerprints of type t . Do you believe him to be the miscreant on this evidence alone? In what size of city would you be convinced?
- 21** Initially urn I contains n red balls and urn II contains n blue balls. A ball selected randomly from urn I is placed in urn II, and a ball selected randomly from urn II is placed in urn I. This whole operation is repeated indefinitely. Given that r of the n balls in urn I are red, find the mass function of R , the number of red balls in urn I after the next repetition.
- Show that the mean of this is $r + 1 - 2r/n$, and hence find the expected number of red balls in urn I in the long run.
- 22** A monkey has a bag with four apples, three bananas, and two pears. He eats fruit at random until he takes a fruit of a kind he has eaten already. He throws that away and the bag with the rest. What is the mass function of the number of fruit eaten, and what is its expectation?
- 23** **Matching** Consider the matching problem of Example 3.17. Let $\mu^{(k)}$ be the k th factorial moment of the number X of matching letters, $\mu^{(k)} = \mathbf{E}(X(X-1)\dots(X-k+1))$. Show that

$$\mu^{(k)} = \begin{cases} 1; & k \leq n \\ 0; & k > n. \end{cases}$$

- 24** Suppose an urn contains m balls which bear the numbers from 1 to m inclusive. Two balls are removed with replacement. Let X be the difference between the two numbers they bear.
- (a) Find $\mathbf{P}(X \leq n)$.
- (b) Show that if $n/m = x$ is fixed as $m \rightarrow \infty$, then $\mathbf{P}(|X| \leq n) \rightarrow 1 - (1-x)^2$; $0 \leq x \leq 1$.
- (c) Show that $\mathbf{E}|X|/m \rightarrow \frac{1}{3}$.
- 25** Suppose the probability of an insect laying n eggs is given by the Poisson distribution with mean $\mu > 0$, that is by the probability distribution over all the nonnegative integers defined by $p_n = e^{-\mu} \mu^n / n!$ ($n = 0, 1, 2, \dots$), and suppose further that the probability of an egg developing is p . Assuming mutual independence of the eggs, show that the probability distribution q_m for the probability that there are m survivors is of the Poisson type and find the mean.
- 26** Preparatory to a camping trip, you can buy six cans of food, all the same size, two each of meat, vegetables, and fruit. Assuming that cans with the same contents have indistinguishable labels, in how many distinguishable ways can the cans be arranged in a row?
- On the trip, there is heavy rain and all the labels are washed off. Show that if you open three of the cans at random the chance that you will open one of each type is $\frac{2}{5}$. If you do not succeed, you continue opening cans until you have one of each type; what is the expected number of open cans?
- 27** A belt conveys tomatoes to be packed. Each tomato is defective with probability p , independently of the others. Each is inspected with probability r ; inspections are also mutually independent. If a tomato is defective and inspected, it is rejected.
- (a) Find the probability that the n th tomato is the k th defective tomato.
- (b) Find the probability that the n th tomato is the k th rejected tomato.
- (c) Given that the $(n+1)$ th tomato is the first to be rejected, let X be the number of its predecessors that were defective. Find $\mathbf{P}(X = k)$, the probability that X takes the value k , and $\mathbf{E}(X)$.
- 28** Mr. Smith must site his widget warehouse in either Acester or Beeley. Initially, he assesses the probability as p that the demand for widgets is greater in Acester, and as $1-p$ that it is greater in Beeley. The ideal decision is to site the warehouse in the town with the larger demand. The cost of the wrong decision, because of increased transport costs, may be assumed to be £1000 if Acester is the correct choice and £2000 if Beeley is the correct choice. Find the expectations of these costs

for each of the two possible decisions, and the values of p for which Acester should be chosen on the basis of minimum expected cost.

Mr. Smith could commission a market survey to assess the demand. If Acester has the higher demand, the survey will indicate this with probability $\frac{3}{4}$ and will indicate Beeley with probability $\frac{1}{4}$. If Beeley has the higher demand the survey will indicate this with probability $\frac{2}{3}$ and will indicate Acester with probability $\frac{1}{3}$. Show the probability that the demand is higher in Acester is $9p/(4 + 5p)$ if the survey indicates Acester. Find also the expected cost for each of the two possible decisions if the survey indicates Acester.

If the survey indicates Acester and $p < 8/17$, where should Mr. Smith site the warehouse?

- 29** A coin is tossed repeatedly and, on each occasion, the probability of obtaining a head is p and the probability of obtaining a tail is $1 - p$ ($0 < p < 1$).
- What is the probability of not obtaining a tail in the first n tosses?
 - What is the probability p_n of obtaining the first tail at the n th toss?
 - What is the expected number of tosses required to obtain the first tail?
- 30** The probability of a day being fine is p if the previous day was fine and is p' if the previous day was wet. Show that, in a consecutive sequence of days, the probability u_n that the n th is fine satisfies $u_n = (p - p')u_{n-1} + p'$, $n \geq 2$.

Show that as $n \rightarrow \infty$, $u_n \rightarrow p'(1 - p + p')^{-1}$.

By considering the alternative possibilities for tomorrow's weather, or otherwise, show that if today is fine the expected number of future days up to and including the next wet day is $1/(1 - p)$.

Show that (today being fine) the expected number of future days up to and including the next two consecutive wet days is $(2 - p)/((1 - p)(1 - p'))$.

- 31** Cars are parked in a line in a parking lot in order of arrival and left there. There are two types of cars, small ones requiring only one unit of parking length (say 15 ft) and large ones requiring two units of parking length (say 30 ft). The probability that a large car turns up to park is p and the probability that a small car turns up is $q = 1 - p$. It is required to find the expected maximum number of cars that can park in a parking length of n units, where n is an integer. Denoting this number by $M(n)$ show that:
- $M(0) = 0$
 - $M(1) = 1 - p$
 - $M(n) - qM(n - 1) - pM(n - 2) = 1$, ($n \geq 2$)

Show that the equations are satisfied by a solution of the form $M(n) = A\alpha^n + B\beta^n + Cn$, where α, β are the roots of the equation $x^2 - qx - p = 0$, and A, B, C are constants to be found. What happens to $M(n)$ as $n \rightarrow \infty$?

- 32** The probability that the postman delivers at least one letter to my house on any day (including Sundays) is p . Today is Sunday, the postman has passed my house and no letter has been delivered.
- What is the probability that at least one letter will be delivered during the next week (including next Sunday)?
 - Given that at least one letter is delivered during the next week, let X be the number of days until the first is delivered. What is $f_X(x)$?
 - What is the expected value of X ?
 - Suppose that all the conditions in the first paragraph hold, except that it is known that a letter will arrive on Thursday. What is the expected number of days until a letter arrives?

- 33** A gambler plays two games, in each of which the probability of her winning is 0.4. If she loses a game she loses her stake, but if she wins she gets double her stake. Suppose that she stakes a in the first game and b in the second, with $a + b = 1$. Show that her expected loss after both games is 0.2.

Suppose she plays again, but now the stake in the first game buys knowledge of the second, so that the chance of winning in the second is ap (≤ 1). Show that the value of a which gives the greatest expected gain is $0.5 + 0.2/p$.

- 34** Let $f_1(X)$ and $f_2(X)$ be functions of the random variable X . Show that (when both sides exist) $[\mathbf{E}(f_1 f_2)]^2 \leq \mathbf{E}(f_1^2)\mathbf{E}(f_2^2)$. Deduce that $\mathbf{P}(X = 0) \leq 1 - [\mathbf{E}(X)]^2/\mathbf{E}(X^2)$.
(Recall that $at^2 + 2bt + c$ has distinct real roots if and only if $b^2 > ac$.)
- 35** Any oyster contains a pearl with probability p independently of its fellows. You have a tiara that requires k pearls and are opening a sequence of oysters until you find exactly k pearls. Let X be the number of oysters you have opened that contain no pearl.
- (a) Find $\mathbf{P}(X = r)$ and show that $\sum_r \mathbf{P}(X = r) = 1$.
 (b) Find the mean and variance of X .
 (c) If $p = 1 - \lambda/k$, find the limit of the distribution of X as $k \rightarrow \infty$.

- 36** A factory produces 100 zoggles a day. Each is defective independently with probability p . If a defective zoggle is sold, it costs the factory £100 in fines and replacement charges. Therefore, each day 10 are selected at random and tested. If they all pass, all 100 zoggles are sold. If more than one is defective, then all 100 zoggles are scrapped. If one is defective, it is scrapped and a further sample of size 10 is taken. If any are defective, the day's output is scrapped; otherwise, 99 zoggles are sold.
- (a) Show that the probability r of not scrapping the day's output is $(1 - p)^{10}(1 + 10p(1 - p)^9)$.
 (b) If testing one zoggle costs £10, find the expected cost of a day's testing.
 (c) Find the expected returns on a day's output in terms of the profit b of a sold zoggle and cost c of a scrapped zoggle.

- 37** An urn contains two blue balls and $n - 2$ red balls; they are removed without replacement.
- (a) Show that the probability of removing exactly one blue ball in $r - 1$ removals is

$$\frac{2(r - 1)(n - r + 1)}{n(n - 1)}.$$

- (b) Show that the probability that the urn first contains no blue balls after the r th removal is

$$\frac{2(r - 1)}{n(n - 1)}.$$

- (c) Find the expected number of removals required to remove both blue balls.

- 38** Suppose that n dice are rolled once; let X be the number of sixes shown. These X dice are rolled again, let Y be the number of sixes shown after this second set of rolls.
- (a) Find the distribution and mean of Y .
 (b) Given that the second set of rolls yielded r sixes, find the distribution and mean of X .
- 39** Pascal and Brianchon now play a series of games that may be drawn (i.e., tied) with probability r . Otherwise, Pascal wins with probability p or Brianchon wins with probability q , where $p + q + r = 1$.
- (a) Find the expected duration of the match if they stop when one or other wins two consecutive games. Also, find the probability that Pascal wins.
 (b) Find the expected duration of the match if they stop when one or other wins two successive games of the games that are won. (That is, draws are counted but ignored.) Find the probability that Pascal wins.

If you were Brianchon and $p > q$, which rules would you rather play by?

- 40** Let the random variable X have a geometric distribution, $\mathbf{P}(X = k) = q^{k-1}p$; $k \geq 1$. Show that for $t > 0$, $\mathbf{P}(X \geq a + 1) \leq pe^{-ta}(1 - qe^t)^{-1}$. Deduce that $\mathbf{P}(X \geq a + 1) \leq (a + 1)p[q(a + 1)a^{-1}]^a$, and compare this with the exact value of $\mathbf{P}(X \geq a + 1)$.
- 41** An archer shoots arrows at a circular target of radius 1 where the central portion of the target inside radius $\frac{1}{4}$ is called the bull. The archer is as likely to miss the target as she is to hit it. When the

archer does hit the target, she is as likely to hit any one point on the target as any other. What is the probability that the archer will hit the bull? What is the probability that the archer will hit k bulls in n attempts? Prove that the mean number of bulls that the archer hits in n attempts is $n/32$.

Show that if the archer shoots 96 arrows in a day, the probability of her hitting no more than one bull is approximately $4e^{-3}$. Show that the average number of bulls the archer hits in a day is 3, and that the variance is approximately $(63\sqrt{3}/64)^2$.

- 42 Prove Chebyshev's inequality that, for a random variable X with mean μ and variance σ^2 ,

$$\mathbf{P}(|X - \mu| \leq h\sigma) \geq 1 - \frac{1}{h^2}, \quad \text{for any } h > 0.$$

When an unbiased coin is tossed n times, let the number of tails obtained be m . Show that

$$\mathbf{P}\left(0.4 \leq \frac{m}{n} \leq 0.6\right) \geq 0.75$$

when $n \geq 100$. Given that $n = 100$, show that

$$\mathbf{P}\left(0.49 \leq \frac{m}{n} \leq 0.51\right) \simeq 3(5\sqrt{(2\pi)})^{-1}.$$

(You may assume Stirling's formula that $n! \simeq \sqrt{(2\pi)n}n^{n+1/2}e^{-n}$ when n is large.)

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

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

$$p_h = \binom{2n-h}{n} \frac{1}{2^{2n-h}},$$

and find the expected value of H , by considering $\sum_{h=0}^n (n-h)p_h$, or otherwise.

- 44 You insure your car. You make a claim in any year with probability q independently of events in other years. The premium in year j is a_j (where $a_j < a_k$ for $k < j$), so long as no claim is made. If you make a claim in year k , then the premium in year $k+j$ is a_j as long as no further claim is made, and so on. Find the expected total payment of premiums until the first claim.
- 45 A Scotch die has faces bearing tartan patterns: three are McDiarmid, two are Meldrum, and one is Murray. Show that the expected number of times you must roll the die before all three patterns have appeared is 7.3.
- 46 **Tail Sums** Let $X \geq 0$ be integer valued. Use the indicator $I(X > k)$ to prove that

$$\mathbf{E}X = \sum_{k \geq 0} \mathbf{P}(X > k),$$

and

$$\mathbf{E}X^r = \sum_{k \geq 0} r k^{r-1} \mathbf{P}(X > k).$$

- 47 **Coupon Collecting; Example (4.3.15) Revisited** Let X_n be the number of coupons collected until you first obtain a coupon that is a duplicate of one you already possess. Find $\mathbf{P}(X_n = k)$ and deduce that

$$(a) \sum_{k=2}^{n+1} \frac{n!}{(n-k+1)!} \frac{k-1}{n^k} = 1.$$

$$(b) \mathbf{E}X_n = \sum_{k=0}^n \frac{n!}{(n-k)!} n^{-k}.$$

48 Let $(x_i; 1 \leq i \leq n)$ be a collection of positive numbers. Show that

$$\left(\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}\right)^{-1} \leq \left(\prod_{i=1}^n x_i\right)^{1/n}.$$

49 If $(y_i; 1 \leq i \leq n)$ is any ordering of $(x_i; 1 \leq i \leq n)$, show that $\sum_{i=1}^n \frac{x_i}{y_i} \geq n$.

50 Let X have finite variance, and set $\nu(x) = \mathbf{E}(X - x)^2$. Show that $\mathbf{E}\nu(X) = 2\text{var}X$.

51 Let X have mean μ , variance σ^2 , and median m . Use (4.6.4) to show that $|\mu - m| < \sigma$.

5

Random Vectors: Independence and Dependence

Wherever there is wealth there will be dependence and expectation.
Samuel Johnson [The Rambler, 189]

5.1 Joint Distributions

Commonly, each outcome of an experiment generates two (or more) real numbers of interest. We can treat these as individual random variables ($X_i; 1 \leq i \leq n$), but it is often important to consider their joint behaviour. For example, if the experiment is your visit to your doctor, you may find out your height H and weight W . These are separate random variables, but are often informative when considered jointly. Thus, the outcome $H = 150$ cm and $W = 150$ kg might disturb your physician, whereas the outcome $H = 190$ cm and $W = 80$ kg probably would not. Likewise, the random vector comprising height, weight, age, sex, blood pressure, and heart rate is of more use considered jointly than separately.

As another example, complicated systems (e.g., space shuttles) have several on-board computers that work together to run the system. If one fails or makes an error, the others can override it; thus, the system fails only when a majority of the computers fail. If X_i is the time until the i th processor fails, then the time until the system fails depends jointly on the collection of random variables, X_1, \dots, X_n . It is natural to refer to such a collection as a random vector, and write $X = (X_1, X_2, \dots, X_n)$. Formally, as before, we have $X = X(\omega); \omega \in \Omega$ and $A_x = \{\omega : X(\omega) = x\} \in \mathcal{F}$, but we do not often refer to the underlying sample space Ω . Because X maps Ω into a countable subset S of R^n , we think of S as the sample space. (You may well have already been doing this instinctively in Chapter 4.)

For simplicity, we summarize the properties of random vectors in two dimensions; the appropriate generalizations in more dimensions are straightforward.

Definition *Let X and Y be two discrete random variables taking values $(x_i; i \geq 1)$ and $(y_j; j \geq 1)$, respectively. Their joint probability mass function $f(x, y)$ is defined by*

$$f(x, y) = \mathbf{P}(X = x, Y = y)$$

as x and y range over all possible values x_i and y_j of X and Y . ▲

The mass function $f(x, y)$ is zero, except at a countable set of points in \mathbb{R}^2 . In fact,

$$(1) \quad f(x, y) \geq 0 \quad \text{for all } x \text{ and } y,$$

and further if $\sum_{i,j} f(x_i, y_j) = 1$, then the joint distribution $f(x, y)$ is not defective. Most important is the **Key Rule**:

$$(2) \quad \mathbf{P}((X, Y) \in C) = \sum_{(x,y) \in C} f(x, y).$$

(3) **Example** Suppose that a coin is tossed twice; let X be the total number of heads shown and Y the total number of tails. Then, (X, Y) takes values in

$$S = \{0, 1, 2\} \times \{0, 1, 2\} = \{(i, j) : i \in \{0, 1, 2\}, j \in \{0, 1, 2\}\}.$$

Clearly, $f(x, y)$ is zero, except at the points $(0, 2)$, $(1, 1)$, and $(2, 0)$. Furthermore,

$$f(0, 2) + f(1, 1) + f(2, 0) = (1 - p)^2 + 2p(1 - p) + p^2 = 1,$$

where we have denoted the probability of a head by p , as usual. ●

Any real function $g(X, Y)$ of two such jointly distributed random variables is itself a random variable. If we set $Z = g(X, Y)$, then Z has a probability mass function given by using the **Key Rule** (2) above:

$$(4) \quad f_Z(z) = \mathbf{P}(g(X, Y) = z) = \sum f(x, y),$$

where the summation is over all x and y , such that $g(x, y) = z$.

In particular, if $g(x, y) = x$, we have

$$(5) \quad f_X(x) = \sum_y f(x, y),$$

and if $g(x, y) = y$, we have

$$(6) \quad f_Y(y) = \sum_x f(x, y).$$

Thus, we have shown the important result that, if we know the joint mass function of several random variables, we can find all their separate mass functions. When obtained in this way, $f_X(x)$ and $f_Y(y)$ are sometimes called *marginal* mass functions.

Here are some examples illustrating joint mass functions.

(7) **Example** A row of n numbered machines are producing components that are identical, except for the serial number. On any day, the k th component produced by the j th machine bears the serial number (j, k) . On the day in question, the r th machine produces c_r ($1 \leq r \leq n$) components, and at the end of the day one component C is picked at random from all those produced. Let its serial number be (X, Y) . Find $f(x, y)$, $f_X(x)$ and $f_Y(y)$.

Solution Because C is picked at random from all $\sum_{r=1}^n c_r$ components, we have

$$f(x, y) = \left(\sum_1^n c_r \right)^{-1}, \quad 1 \leq x \leq n; 1 \leq y \leq c_x, \\ = a \quad (\text{say}).$$

Then, by (5),

$$f_X(x) = \sum_y f(x, y) = ac_x.$$

Now define the function $H(i, j) = \begin{cases} 1 & \text{if } c_i \geq c_j, \\ 0 & \text{otherwise.} \end{cases}$

Then, by (6),

$$f_Y(y) = \sum_x f(x, y) = a \sum_{x=1}^n H(x, y). \quad \bullet$$

(8) Example: Cutting for the Deal It is customary, before engaging in a card game, to cut for the deal; each player removes a portion of the deck in turn, and then each reveals the bottom card of his segment. The highest card wins. For these to be random variables, we need to assign numerical values to the court cards, so we set J = 11, Q = 12, K = 13, A = 14, when aces are high.

- Art and Bart cut for deal, aces high. Let X be Art's card, and Y be Bart's card. Find the joint mass function of X and Y . Does it make any difference how many cards Art removes from the deck?
- Let V be the loser's card, and W the dealer's (winning) card. Find the joint mass function of V and W , and the separate mass functions of V and W .
- Find the mass function of the dealer's winning margin (namely, $W - V$).
- What is the mass function of the dealer's card when three players cut for deal?

Note that in the event of a tie, the deck is shuffled and the players cut again to choose the dealer.

Solution (a) Each random variable takes values in $\{2, 3, \dots, 14\}$. Cutting the deck twice amounts to selecting two cards at random, and because ties are not allowed, $X \neq Y$. By symmetry, any two unequal values are equally likely to occur, so

$$f_{X,Y}(x, y) = \mathbf{P}(X = x, Y = y) = \begin{cases} \frac{1}{12} \cdot \frac{1}{13} & x \neq y \\ 0 & x = y \end{cases} \quad 2 \leq x, y \leq 14.$$

It makes no difference how many cards Art removes with his cut.

(b) Of course, $W > V$, so

$$f_{V,W}(v, w) = \mathbf{P}(V = v, W = w) = \mathbf{P}(X = v, Y = w) + \mathbf{P}(X = w, Y = v) \\ = \begin{cases} \frac{2}{12} \cdot \frac{1}{13} = \frac{1}{78} & 2 \leq v < w \leq 14 \\ 0 & \text{otherwise.} \end{cases}$$

This is otherwise obvious, because the experiment amounts to choosing an unordered pair of unequal cards at random, with equal probability of choosing any pair. Hence, for $v < w$, $f(v, w) = \binom{13}{2}^{-1}$, as above. Now, by (5),

$$f_V(v) = \sum_{w=v+1}^{14} \frac{1}{78} = \frac{14-v}{78}; \quad 2 \leq v \leq 13.$$

Then by (6),

$$f_W(w) = \sum_2^{w-1} \frac{1}{78} = \frac{w-2}{78}; \quad 3 \leq w \leq 14.$$

(c) By (4),

$$f_Z(z) = \mathbf{P}(W - V = z) = \sum \frac{1}{78},$$

where the summation is over all v and w such that $w - v = z$. Because $z \leq v < w \leq 14$, there are exactly $13 - z$ terms in this sum, so

$$\mathbf{P}(W - V = z) = \frac{13-z}{78}; \quad 1 \leq z \leq 12.$$

(d) Arguing as we did for (b), where now $u < v < w$, we have

$$\mathbf{P}(U = u, V = v, W = w) = \binom{13}{3}^{-1}.$$

Hence,

$$\begin{aligned} f_W(w) &= \sum_{2 \leq u < v < w} \binom{13}{3}^{-1}; \quad 4 \leq w \leq 14, \\ &= \sum_{3 \leq v < w} (v-2) \binom{13}{3}^{-1} = \frac{1}{2}(w-3)(w-2) \binom{13}{3}^{-1}. \end{aligned}$$

(9) Example For what value of c is the function

$$f(x, y) = c \binom{x+y-1}{x} \lambda^x \mu^y, \quad x \geq 0; y \geq 1; 1 > 1 - \lambda > \mu > 0$$

a joint mass function? For this value of c , find the mass functions of X and Y .

Solution By (2),

$$\begin{aligned} c^{-1} &= \sum_{y=1}^{\infty} \sum_{x=0}^{\infty} \binom{x+y-1}{x} \lambda^x \mu^y = \sum_{y=1}^{\infty} \frac{\mu^y}{(1-\lambda)^y} \quad \text{by (3.6.12),} \\ &= \frac{\mu}{1-\lambda-\mu}. \end{aligned}$$

Then, by (5),

$$f_X(x) = c \sum_{y=1}^{\infty} \mu \lambda^x \binom{x+1+y-2}{y-1} \mu^{y-1} = \frac{(1-\lambda-\mu)\lambda^x}{(1-\mu)^{x+1}}, \quad x \geq 0.$$

Likewise,

$$f_Y(y) = c \sum_{x=0}^{\infty} \binom{x+y-1}{x} \lambda^x \mu^y = \frac{(1-\lambda-\mu)\mu^{y-1}}{(1-\lambda)^y}, \quad y \geq 1.$$

Thus, $X + 1$ and Y are both geometric, with parameters $\frac{\lambda}{1-\mu}$ and $\frac{\mu}{1-\lambda}$, respectively, X taking values in the nonnegative integers and Y in the positive integers. ●

- (10) **Example** Leif and Rolf are bored with fair games. They want to play a game in which the probability of winning (for Leif) is λ , where λ is an arbitrary number in $[0, 1]$. Also, they want the game to be of finite duration with probability 1. Unfortunately, the only gaming aid they have is a fair coin. Can you supply them with a game?

Solution Let λ have binary expansion

$$\lambda = 0.b_1b_2b_3 \dots = \sum_{n=1}^{\infty} b_n 2^{-n}.$$

Now toss the coin repeatedly and let I_n be the indicator of the event that the n th toss is a head. Let T be the first toss such that $I_n \neq b_n$, $T = \min\{n: I_n \neq b_n\}$. If $I_T < b_T$, then Leif wins; otherwise, Rolf wins. Now, $\mathbf{P}(T = n) = (\frac{1}{2})^n$ so that

$$\mathbf{P}(T < \infty) = \sum_1^{\infty} \left(\frac{1}{2}\right)^n = 1$$

and indeed $\mathbf{E}(T) = 2$. Also, Leif can only win at the n th toss if $b_n = 1$ so

$$\mathbf{P}(\text{Leif wins}) = \sum_n b_n \mathbf{P}(T = n) = \sum_n b_n 2^{-n} = \lambda,$$

as required. ●

5.2 Independence

Given the joint mass function of X and Y , equations (5.1.5) and (5.1.6) yield the marginal mass functions of X and Y . However, to be given the marginal distributions does not in general uniquely determine a joint distribution.

- (1) **Example** Let X and Y have joint mass function given by

$$f(0, 0) = \frac{1}{6}, \quad f(0, 1) = \frac{1}{3}, \quad f(1, 0) = \frac{1}{12}, \quad f(1, 1) = \frac{5}{12},$$

and let U and V have joint mass function given by

$$f(0, 0) = \frac{1}{4}, \quad f(0, 1) = \frac{1}{4}, \quad f(1, 0) = 0, \quad f(1, 1) = \frac{1}{2}.$$

Then, summing to get the marginal mass functions shows that:

$$f_X(0) = f_U(0) = \frac{1}{2}, \quad f_X(1) = f_U(1) = \frac{1}{2};$$

and

$$f_Y(0) = f_V(0) = \frac{1}{4}, \quad f_Y(1) = f_V(1) = \frac{3}{4}.$$

These marginal mass functions are the same, but the joint mass functions are different. ●

There is one exceptionally important special case when marginal mass functions do determine the joint mass function uniquely.

(2) **Definition** Random variables X and Y are independent if, for all x and y ,

$$f(x, y) = f_X(x)f_Y(y).$$

This is equivalent to $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$, where $A = \{\omega: X(\omega) = x\}$ and $B = \{\omega: X(\omega) = y\}$, which is the definition of independence for the events A and B . More generally, a collection $(X_i; 1 \leq i \leq n)$ with mass function f is independent if for all $x = (x_1, \dots, x_n)$

(3)
$$f(x) = \prod_{i=1}^n f_{X_i}(x_i).$$
 ▲

Note that if X or Y (or both) are improper random variables [so that $f(x, y)$ is defective], then to say they are independent is interpreted as meaning

(4)
$$\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x)\mathbf{P}(Y = y)$$

for all finite x and y . This may seem odd, but such random variables occur quite naturally in simple random walks and other topics.

Example 5.1.7 Revisited Recall that n machines produce components. Suppose that all the machines produce the same number c of components, and as before we pick one at random and let its serial number be (X, Y) , where X is the machine number and Y is the component index. Then,

$$f(x, y) = (nc)^{-1}; \quad 1 \leq x \leq n; \quad 1 \leq y \leq c$$

and

$$f_X(x) = \frac{1}{n}; \quad 1 \leq x \leq n \quad \text{and} \quad f_Y(y) = \frac{1}{c}; \quad 1 \leq y \leq c.$$

Obviously, $f(x, y) = f_X(x)f_Y(y)$ and so X and Y are independent in this case. ●

- (5) **Example 1 Revisited** Observe that the mass functions of X and Y , and of U and V , do not satisfy Definition 2. Let W and Z be independent random variables such that $f_W(0) = \frac{1}{2}$, $f_W(1) = \frac{1}{2}$ and $f_Z(0) = \frac{1}{4}$, $f_Z(1) = \frac{3}{4}$.

Then, by Definition 2, their joint mass function is

$$f(0, 0) = \frac{1}{8}, \quad f(0, 1) = \frac{3}{8}, \quad f(1, 0) = \frac{1}{8}, \quad f(1, 1) = \frac{3}{8}. \quad \bullet$$

- (6) **Example 5.1.9 Revisited** Observe that X and Y are not independent because $\binom{x+y-1}{x} \lambda^x \mu^y$ cannot be expressed in the form $f_X(x)f_Y(y)$. If $X+1$ and Y were independent geometric random variables with parameters $\frac{\lambda}{1-\mu}$ and $\frac{\mu}{1-\lambda}$, then the joint mass function would be

$$f(x, y) = \left(1 - \frac{\lambda}{1-\mu}\right) \left(1 - \frac{\mu}{1-\lambda}\right) \left(\frac{\lambda}{1-\mu}\right)^x \left(\frac{\mu}{1-\lambda}\right)^{y-1}, \quad x \geq 0; y \geq 1. \quad \bullet$$

The apparently simple Definition 2 implies a great deal more about independent random variables, as the following result shows.

- (7) **Theorem** *Let X and Y be independent random variables. Then:*

(a) *For arbitrary countable sets A and B ,*

$$(8) \quad \mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B),$$

and

(b) *For any real functions $g(\cdot)$ and $h(\cdot)$, $g(X)$ and $h(Y)$ are independent.*

Proof (a) The left-hand side of (8) is

$$\begin{aligned} \sum_{x \in A} \sum_{y \in B} \mathbf{P}(X = x, Y = y) &= \sum_{x \in A} \sum_{y \in B} \mathbf{P}(X = x)\mathbf{P}(Y = y) \quad \text{by independence} \\ &= \sum_{x \in A} \mathbf{P}(X = x) \sum_{y \in B} \mathbf{P}(Y = y) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B), \end{aligned}$$

as required. For (b), let $A = \{x: g(X) = \xi\}$ and $B = \{y: h(Y) = \eta\}$. Then, by part (a), for any ξ and η ,

$$\begin{aligned} \mathbf{P}(g(X) = \xi, h(Y) = \eta) &= \mathbf{P}(X \in A, Y \in B) = \mathbf{P}(X \in A)\mathbf{P}(Y \in B) \\ &= \mathbf{P}(g(X) = \xi)\mathbf{P}(h(Y) = \eta), \end{aligned}$$

as required. ■

- (9) **Example** Independent random variables X and Y take the values -1 or $+1$ only, and $\mathbf{P}(X = 1) = a$, $\mathbf{P}(Y = 1) = \alpha$. A third random variable Z is defined by $Z = \cos((X + Y)\frac{\pi}{2})$. If $0 < a, \alpha < 1$, show that there are unique values of a and α such that X and Z are independent, and Y and Z are independent. In this case, are X, Y , and Z independent?

Solution First, for Z ,

$$\mathbf{P}(Z = 1) = \mathbf{P}(X + Y = 0) = a(1 - \alpha) + \alpha(1 - a)$$

and, likewise, $\mathbf{P}(Z = -1) = a\alpha + (1 - a)(1 - \alpha)$. Now,

$$\mathbf{P}(Z = 1, X = 1) = \mathbf{P}(X = 1, Y = -1) = a(1 - \alpha).$$

Hence, if

$$(10) \quad a(1 - \alpha) = a(a(1 - \alpha) + \alpha(1 - a))$$

we have $\mathbf{P}(Z = 1, X = 1) = \mathbf{P}(Z = 1)\mathbf{P}(X = 1)$. Simplifying (10) yields $\alpha = \frac{1}{2}$. Now, plodding through three similar constraints shows that X and Z are independent iff $\alpha = a = \frac{1}{2}$. By symmetry, the same condition holds iff Y and Z are independent.

However, X , Y , and Z are not independent because

$$\mathbf{P}(X = 1, Y = 1, Z = -1) = 0 \neq \mathbf{P}(X = 1)\mathbf{P}(Y = 1)\mathbf{P}(Z = -1). \quad \bullet$$

Independent random variables often have interesting and useful properties.

- (11) **Example** Let X and Y be independent geometric random variables having respective mass functions $f_X(x) = (1 - \lambda)\lambda^x$ and $f_Y(y) = (1 - \mu)\mu^y$ for $x \geq 0$ and $y \geq 0$. What is the mass function of $Z = \min\{X, Y\}$?

Solution By independence,

$$\mathbf{P}(Z > n) = \mathbf{P}(X > n \cap Y > n) = \mathbf{P}(X > n)\mathbf{P}(Y > n) = \lambda^{n+1}\mu^{n+1} = (\lambda\mu)^{n+1}.$$

Hence,

$$\mathbf{P}(Z = n) = \mathbf{P}(Z > n - 1) - \mathbf{P}(Z > n) = (1 - \lambda\mu)(\lambda\mu)^n$$

and Z is also geometric with parameter $\lambda\mu$. ●

5.3 Expectation

Let the random variable $Z = g(X, Y)$ be a function of X and Y . Using (5.1.4) and the definition of expectation (4.3.1), we have

$$\mathbf{E}(Z) = \sum_z z f_Z(z) = \sum_z z \mathbf{P}(g(X, Y) = z).$$

This expression for $\mathbf{E}(Z)$ is not always simple or convenient for use in calculation. The following generalization of Theorem 4.3.4 is therefore very useful.

- (1) **Theorem** Let X and Y have joint mass function $f(x, y)$. Whenever the sum on the right-hand side is absolutely convergent, we have

$$\mathbf{E}(g(X, Y)) = \sum_{x,y} g(x, y) f(x, y).$$

Proof The proof is essentially the same as that of Theorem 4.3.4. ■

(2) **Corollary** For any real numbers a and b ,

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$$

when both sides exist and are finite.

Proof Because the sum is absolutely convergent, by (1),

$$\begin{aligned} \mathbf{E}(aX + bY) &= \sum_{x,y} (ax + by)f(x, y) = \sum_{x,y} axf(x, y) + \sum_{x,y} byf(x, y) \\ &= \sum_x axf_X(x) + \sum_y byf_Y(y) && \text{by (5.1.5) and (5.1.6)} \\ &= a\mathbf{E}(X) + b\mathbf{E}(Y). \quad \blacksquare \end{aligned}$$

(3) **Example: Coupons** Recall Example 4.3.15 in which you were collecting coupons; we can now find $\mathbf{E}(R)$ more quickly. Let T_1 be the number of packets required to obtain the first coupon, T_2 the *further* number of packets required to obtain a second type of coupon, T_3 the *further* number required for the third type and so on. Then,

$$R = \sum_{k=1}^n T_k.$$

Obviously, $T_1 = 1$. Also,

$$\mathbf{P}(T_2 = r) = \left(\frac{1}{n}\right)^{r-1} \left(1 - \frac{1}{n}\right)$$

so that T_2 is geometric with mean $\frac{n}{n-1}$. Likewise, T_k is geometric with mean

$$\mathbf{E}(T_k) = \frac{n}{n - k + 1}; \quad 1 \leq k \leq n.$$

Hence, by (2),

$$\mathbf{E}(R) = \sum_{k=1}^n \mathbf{E}(T_k) = \sum_{k=1}^n \frac{n}{n - k + 1},$$

the same as the answer obtained with somewhat more effort in Example 4.3.15. \bullet

Corollary (2) is often useful when considering sums of indicators. For example, let $\{A_1, A_2, \dots, A_n\}$ be any collection of events, and let

$$I_i = \begin{cases} 1 & \text{if } A_i \text{ occurs} \\ 0 & \text{if } A_i \text{ does not occur} \end{cases}$$

be the indicator of A_i . Now, let X be the number of the A_i that occur. Then, by construction $X = \sum_{i=1}^n I_i$, and by (2)

$$\mathbf{E}(X) = \sum_{i=1}^n \mathbf{E}(I_i) = \sum_{i=1}^n \mathbf{P}(A_i).$$

We use this result in the following example.

Example: Binomial Distribution Let X be binomial with parameters n and p . Show that the factorial moments of X are given by

$$\mu^{(k)} = p^k n(n-1)\dots(n-k+1).$$

Solution Suppose a coin that shows heads with probability p is tossed n times. Then, the number of heads has the mass function of X . Let Y be the number of distinct sets of k such that all k tosses show heads. Then, $Y = \binom{X}{k}$. However, each of the $\binom{n}{k}$ distinct sets of k tosses shows k heads with probability p^k . Hence, $\mathbf{E}(Y) = \binom{n}{k} p^k$. Therefore, we have

$$\mathbf{E}\left(\binom{X}{k}\right) = \binom{n}{k} p^k,$$

which is the desired result. ●

We single certain expectations out for special notice. Just as random variables have moments, jointly distributed random variables have joint moments.

(4) **Definition** The joint moments of X and Y are

$$\mu_{ij} = \mathbf{E}(X^i Y^j); \quad i, j \geq 1. \quad \blacktriangle$$

(5) **Definition** The covariance of X and Y is

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))] = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y). \quad \blacktriangle$$

This is the most important of the *central joint moments*, which are

$$\sigma_{ij} = \mathbf{E}[(X - \mathbf{E}(X))^i (Y - \mathbf{E}(Y))^j]; \quad i, j \geq 1.$$

Here are two interesting properties of $\text{cov}(X, Y)$.

(6) **Theorem** For jointly distributed random variables X and Y , and constants a, b, c, d , we have:

- (i) $\text{cov}(aX + b, cY + d) = ac \text{cov}(X, Y)$
- (ii) $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y)$

Proof

$$\begin{aligned} \text{(i) } \text{cov}(aX + b, cY + d) &= \mathbf{E}[(aX + b - a\mathbf{E}(X) - b)(cY + d - c\mathbf{E}(Y) - d)] \\ &= \mathbf{E}[ac(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))] = ac \text{cov}(X, Y) \end{aligned}$$

$$\begin{aligned} \text{(ii) } \text{var}(X + Y) &= \mathbf{E}(X + Y - \mathbf{E}(X) - \mathbf{E}(Y))^2 \\ &= \mathbf{E}[(X - \mathbf{E}(X))^2 + (Y - \mathbf{E}(Y))^2 + 2(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))], \end{aligned}$$

as required. ■

Let us find $\text{cov}(X, Y)$ for the simple examples we have met above.

- (1) **Examples 5.2.1 and 5.2.5 Revisited** Find the covariance for each of the three joint mass functions given in these two examples.

Solution In every case,

$$\mathbf{E}(XY) = 1^2 f(1, 1)$$

and $\mathbf{E}(X) = f(1, 0) + f(1, 1)$, and $\mathbf{E}(Y) = f(0, 1) + f(1, 1)$.

Hence,

$$\text{cov}(X, Y) = f(1, 1) - (f(1, 0) + f(1, 1))(f(0, 1) + f(1, 1)).$$

Evaluating this in the three given instances shows that:

- (i) $\text{cov}(X, Y) = \frac{5}{12} - \left(\frac{1}{3} + \frac{5}{12}\right)\left(\frac{1}{12} + \frac{5}{12}\right) = \frac{1}{24}$
- (ii) $\text{cov}(U, V) = \frac{1}{2} - \left(\frac{1}{4} + \frac{1}{2}\right) \cdot \frac{1}{2} = \frac{1}{8}$
- (iii) $\text{cov}(W, Z) = \frac{3}{8} - \left(\frac{3}{8} + \frac{3}{8}\right)\left(\frac{1}{8} + \frac{3}{8}\right) = 0.$ ●

- (7) **Example 5.1.8 Revisited** Recall that Art and Bart are cutting for the deal. Find $\text{cov}(X, Y)$ and $\text{cov}(V, W)$.

Solution

$$\mathbf{E}(X) = \mathbf{E}(Y) = \sum_2^{14} \frac{x}{13} = 8.$$

Also, using (1),

$$\begin{aligned} \mathbf{E}(XY) &= \sum_{2 \leq x \neq y \leq 14} \frac{x}{12} \cdot \frac{y}{13} = \frac{1}{12} \cdot \frac{1}{13} \sum_y (105 - y - 1)y \\ &= \frac{1}{12} \cdot \frac{1}{13} (105 \times 104 - 1118) = 64 - \frac{7}{6}. \end{aligned}$$

Hence, $\text{cov}(X, Y) = -\frac{7}{6}$.

Likewise, using the expressions in Example 5.1.8 for the marginal mass functions of V and W , we find

$$\begin{aligned} \mathbf{E}(V) &= \sum_{v=2}^{13} \frac{v(14-v)}{78} = \frac{1}{78} \sum_{v=2}^{13} (13v - v(v-1)) \\ &= +\frac{1}{78} \sum_{v=2}^{13} \left(\frac{13}{2} [v(v+1) - v(v-1)] \right. \\ &\quad \left. - \frac{1}{3} [(v+1)v(v-1) - v(v-1)(v-2)] \right) \\ &= \frac{17}{3} \end{aligned}$$

after successive cancellation of the terms in the sum.

Similarly, we find $\mathbf{E}(W) = \frac{31}{3}$.

Now for all ω , $X(\omega)Y(\omega) = V(\omega)W(\omega)$, so $\mathbf{E}(VW) = \mathbf{E}(XY)$, and finally,

$$\text{cov}(V, W) = 62 + \frac{5}{6} - \frac{17}{3} \cdot \frac{31}{3} = +\frac{77}{18}. \quad \bullet$$

Just as joint mass functions have a simple form when random variables are independent, so too do joint moments simplify.

- (8) **Theorem** *If X and Y are independent random variables with finite expectations, then $\mathbf{E}(XY)$ exists, and $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. It follows that $\text{cov}(X, Y) = 0$ in this case.*

Proof By independence and Theorem 5.3.1,

$$\mathbf{E}(|XY|) = \sum_{x,y} |xy| f_X(x) f_Y(y) = \sum_x |x| f_X(x) \sum_y |y| f_Y(y) = \mathbf{E}(|X|)\mathbf{E}(|Y|)$$

so $\mathbf{E}(|XY|) < \infty$. Thus, $\mathbf{E}(XY)$ exists, and the same argument shows that $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. ■

- (9) **Definition** *If $\text{cov}(X, Y) = 0$, then X and Y are said to be uncorrelated. If $\mathbf{E}(XY) = 0$, then X and Y are said to be orthogonal.* ▲

It follows that independent random variables are uncorrelated, but the converse is not true, as the following example shows.

- (10) **Example** A random variable X is said to be *symmetric* if $\mathbf{P}(X = -x) = \mathbf{P}(X = x)$ for all x . Let X be symmetric with $\mathbf{E}(X^3) < \infty$, and let $Y = X^2$. Then, because X has an expectation it is zero, by symmetry, and

$$\mathbf{E}(XY) = \mathbf{E}(Y^3) = \sum_{x>0} x^3 (f(x) - f(-x)) = 0 = \mathbf{E}(X)\mathbf{E}(Y).$$

Thus, $\text{cov}(X, Y) = 0$, even though X and Y are not independent. In this case, X and Y are uncorrelated and orthogonal, but dependent. ●

Thus, up to a point, and in a way that we carefully leave unspecified, $\text{cov}(X, Y)$ can be an indication of the dependence of X and Y . It has the drawback that it depends on the scale of X and Y . Thus, if a is a constant, aX and Y have the same “dependence” as X and Y

(whatever we mean by that), but $\text{cov}(aX, Y) = a \text{cov}(X, Y)$. For this reason, statisticians more commonly use the following.

(11) **Definition** *The correlation coefficient of random variables X and Y is*

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{(\text{var}(X) \text{var}(Y))^{\frac{1}{2}}},$$

whenever the right-hand side exists. ▲

Example 5.1.3 Revisited Here, X and Y are the number of heads and tails, respectively, when a coin is tossed twice. What are $\text{cov}(X, Y)$ and $\rho(X, Y)$?

Solution Trivially,

$$\mathbf{E}(XY) = 1^2 \mathbf{P}(X = 1, Y = 1) = 2p(1 - p).$$

Likewise, $\mathbf{E}(X) = 2p$, $\mathbf{E}(Y) = 2(1 - p)$, $\text{var}(X) = 2p(1 - p)$ and $\text{var}(Y) = 2p(1 - p)$. Hence,

$$\text{cov}(X, Y) = 2p(1 - p) - 4p(1 - p) = -2p(1 - p)$$

and

$$\rho(X, Y) = \frac{-2p(1 - p)}{(4p^2(1 - p)^2)^{\frac{1}{2}}} = -1. \quad \bullet$$

The correlation coefficient ρ has the following interesting properties; we assume that X and Y are not constant, and have finite variance.

(12) **Theorem** *If X and Y have correlation $\rho(X, Y)$, then:*

- (i) $-1 \leq \rho(X, Y) \leq 1$.
- (ii) $|\rho| = 1$ if and only if $\mathbf{P}(X = aY) = 1$ for some constant a .
- (iii) $\rho(aX + b, cY + d) = \text{sgn}(ac)\rho(X, Y)$, where $\text{sgn}(x)$ denotes the sign of x .
- (iv) $\rho = 0$ if X and Y are independent.

The proof of this theorem relies on the following important and useful result.

(13) **Lemma: Cauchy-Schwarz Inequality** *If $\mathbf{E}(X^2)\mathbf{E}(Y^2) < \infty$, then*

$$(14) \quad (\mathbf{E}(XY))^2 \leq \mathbf{E}(X^2)\mathbf{E}(Y^2).$$

Proof Suppose $0 < \mathbf{E}(X^2)\mathbf{E}(Y^2)$. By Theorem 4.3.6 (iii),

$$(15) \quad \begin{aligned} 0 &\leq \mathbf{E}[(X\mathbf{E}(Y^2) - Y\mathbf{E}(XY))^2] \\ &= \mathbf{E}(X^2)(\mathbf{E}(Y^2))^2 - 2\mathbf{E}(XY)^2\mathbf{E}(Y^2) + \mathbf{E}(Y^2)[\mathbf{E}(XY)]^2 \\ &= \mathbf{E}(Y^2)[\mathbf{E}(X^2)\mathbf{E}(Y^2) - (\mathbf{E}(XY))^2]. \end{aligned}$$

Because $\mathbf{E}(Y^2) > 0$, (14) follows. Of course, (14) is trivially true if $\mathbf{E}(Y^2) = 0$, for then $Y = XY = 0$ with probability one. ■

Proof of (12) (i) Applying Lemma 13 to the random variables $X - \mathbf{E}(X)$ and $Y - \mathbf{E}(Y)$ shows that $(\rho(X, Y))^2 \leq 1$, and so $-1 \leq \rho \leq 1$, as required.

(ii) If $|\rho| = 1$, then from (15),

$$\mathbf{E}[(X\mathbf{E}(Y^2) - Y\mathbf{E}(XY))^2] = 0,$$

and so from Example 4.6.10, with probability one $X = (\mathbf{E}(XY)/\mathbf{E}(Y^2))Y$.

(iii) Expanding, and using Theorem 6(i),

$$\rho(aX + b, cY + d) = \frac{ac \operatorname{cov}(X, Y)}{(a^2 \operatorname{var}(X)c^2 \operatorname{var}(Y))^{\frac{1}{2}}} = \frac{ac}{\sqrt{(ac)^2}} \rho(X, Y),$$

as required

(iv) This follows immediately from Theorem 8. ■

(16) Example (5.1.9) Revisited Recall that X and Y have joint mass function

$$f(x, y) = \frac{1 - \lambda - \mu}{\mu} \binom{x + y - 1}{x} \lambda^x \mu^y.$$

Show that $\rho(X, Y) = \left(\frac{\lambda\mu}{(1-\lambda)(1-\mu)} \right)^{\frac{1}{2}}$.

Solution First, we calculate $\mathbf{E}(XY)$ as

$$\begin{aligned} \frac{1 - \lambda - \mu}{\mu} \sum_{x,y} xy \binom{x + y - 1}{x} \lambda^x \mu^y &= \frac{1 - \lambda - \mu}{\mu} \sum_{y=1}^{\infty} y^2 \mu^y \lambda \\ &\times \sum_{x=1}^{\infty} \binom{x + y - 1}{x - 1} \lambda^{x-1} = \frac{(1 - \lambda - \mu)\lambda}{(1 - \lambda)\mu} \sum_{y=1}^{\infty} y^2 \left(\frac{\mu}{1 - \lambda} \right)^y \\ &= \frac{\lambda(1 - \lambda + \mu)}{(1 - \lambda - \mu)^2}. \end{aligned}$$

Now we have already discovered in Example 5.1.9 that X and Y have geometric mass functions, so by Example 4.3.13

$$\begin{aligned} \mathbf{E}(X) &= \frac{1 - \mu}{1 - \lambda - \mu} - 1, & \mathbf{E}(Y) &= \frac{1 - \lambda}{1 - \lambda - \mu} \\ \operatorname{var}(X) &= \frac{\mu(1 - \lambda)}{(1 - \lambda - \mu)^2}, & \operatorname{var}(Y) &= \frac{\lambda(1 - \mu)}{(1 - \lambda - \mu)^2} \end{aligned}$$

and plugging all this into (11) yields

$$\rho = \frac{\lambda(1 - \lambda + \mu) - \lambda(1 - \lambda)}{(\mu(1 - \lambda)\lambda(1 - \mu))^{\frac{1}{2}}},$$

as required. ●

Finally, we remark that $\text{cov}(X, Y)$ and $\rho(X, Y)$ are not the only functions used to measure dependence between X and Y . Another such function is

$$(17) \quad I(X, Y) = \sum_x \sum_y f(x, y) \log \left(\frac{f(x, y)}{f_X(x) f_Y(y)} \right) = \mathbf{E} \left(\log \left(\frac{f(X, Y)}{f_X(X) f_Y(Y)} \right) \right).$$

See Example 5.16 for more on this.

5.4 Sums and Products of Random Variables: Inequalities

These arise in many ways. For example, it is often useful to write a random variable as a sum of simpler random variables.

- (1) **Example: Binomial Random Variable** The random variable X with mass function

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

has arisen in many ways; classically, it is the number of heads in n tosses of a biased coin. We now see that we can think about X in a different way. Let I_k be the indicator of the event that the k th toss of the coin shows a head. Then,

$$X = I_1 + I_2 + \cdots + I_n = \sum_{k=1}^n I_k.$$

We have written X as a sum of Bernoulli trials or indicators. Hence,

$$\mathbf{E}(X) = \mathbf{E} \left(\sum_{k=1}^n I_k \right) = \sum_{k=1}^n \mathbf{E}(I_k) = np.$$

Likewise,

$$\mathbf{E}(X^2) = \mathbf{E} \left(\sum_{k=1}^n I_k \right)^2 = \sum_{k=1}^n \mathbf{E}(I_k^2) + \sum_{j \neq k} \mathbf{E}(I_j I_k) = np + n(n-1)p^2.$$

Hence, $\text{var}(X) = np(1-p)$. You should compare this with your earlier methods using

$$\mathbf{E}(X^2) = \sum_{k=1}^n k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^n (k(k-1) + k) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

and so on. ●

- (2) **Theorem** Any discrete random variable X can be written as a linear combination of indicator random variables; thus,

$$X = \sum_i a_i I(A_i)$$

for some collection of events $(A_i; i \geq 1)$ and real numbers $(a_i; i \geq 1)$.

Proof Just let $(a_i; i \geq 1)$ include the set of possible values of X , and set $A_i = \{\omega: X(\omega) = a_i\}$. ■

- (3) **Example: Matching** Suppose that n distinct numbered keys ordinarily hang on n hooks bearing corresponding distinct numbers. On one occasion an inebriated turnkey hangs the keys at random on the hooks (one to each hook). Let X be the number of keys, which are then on the correct hooks. Find $\mathbf{E}(X)$ and $\text{var}(X)$.

Solution Let I_j be the indicator of the event that the j th key does hang on the j th hook. Then,

$$X = \sum_{j=1}^n I_j.$$

Now by symmetry $\mathbf{P}(I_j = 1) = 1/n$ and for $j \neq k$,

$$(4) \quad \mathbf{P}(I_j I_k = 1) = \frac{1}{n(n-1)}.$$

Hence,

$$\mathbf{E}(X) = \mathbf{E}\left(\sum_{j=1}^n I_j\right) = \sum_{j=1}^n \mathbf{E}(I_j) = \sum_{j=1}^n \mathbf{P}(I_j = 1) = 1.$$

Also,

$$\mathbf{E}(X^2) = \mathbf{E}\left(\sum_{j=1}^n I_j^2 + \sum_{j \neq k} I_j I_k\right) = 1 + 1,$$

using (4), and the fact that $I_j^2 = I_j$. Hence, $\text{var}(X) = 1$. ●

Indicators can also be useful when multiplied together; here is an illustration.

- (5) **Example** Let us prove (1.4.8). Recall that we have events A_1, \dots, A_n , and we seek the probability that at least one of them occurs, namely,

$$\mathbf{P}\left(\bigcup_{j=1}^n A_j\right) = t_n \quad (\text{say}).$$

For economy of notation, we set

$$(6) \quad s_r = \sum_{i_1 < i_2 < \dots < i_r} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_r}); \quad 1 \leq r \leq n.$$

We let I_j be the indicator of the event that A_j occurs, and set

$$S_r = \sum_{i_1 < i_2 < \dots < i_r} I_{i_1} I_{i_2} \dots I_{i_r}.$$

Next, observe that

$$(7) \quad 1 - \prod_{j=1}^n (1 - I_j) = \begin{cases} 1 & \text{if at least one } A_j \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Hence, this is the indicator of the event whose probability we seek, and

$$\begin{aligned} \mathbf{P}\left(\bigcup_{j=1}^n A_j\right) &= \mathbf{E}\left(1 - \prod_{j=1}^n (1 - I_j)\right) \\ &= \mathbf{E}(S_1 - S_2 + \cdots + (-)^{n+1} S_n) \quad \text{on multiplying out,} \\ &= s_1 - s_2 + \cdots + (-)^{n+1} s_n \quad \text{by (6), as required.} \quad \bullet \end{aligned}$$

The same expansion (7) can be used to prove the following interesting inequalities.

Theorem **Theorem: Inclusion–Exclusion Inequalities** *With the notation of Example 5, for $1 \leq r \leq n$,*

$$(-)^r \left(\mathbf{P}\left(\bigcup_{j=1}^n A_j\right) - s_1 + s_2 - \cdots + (-)^r s_r \right) \geq 0.$$

Proof First, we prove a simple identity. Obviously, $(1+x)^k(1+\frac{1}{x})^{-1} = x(1+x)^{k-1}$. Hence, equating the coefficient of x^{r+1} on each side gives

$$(8) \quad \binom{k}{r+1} - \binom{k}{r+2} + \cdots + (-)^{k-r+1} \binom{k}{k} = \binom{k-1}{r} \geq 0.$$

Furthermore, we have

$$(9) \quad \begin{aligned} 1 - \prod_{j=1}^n (1 - I_j) - S_1 + S_2 - \cdots + (-)^r S_r \\ = (-)^r (S_{r+1} - S_{r+2} + \cdots + (-)^{n-r+1} S_n). \end{aligned}$$

Now suppose exactly k of A_1, \dots, A_n occur. If $k \leq r$, the right-hand side of (9) is zero. If $k > r$, the contribution in the bracket on the right-hand side is

$$\binom{k}{r+1} - \binom{k}{r+2} + \cdots + (-)^{k-r+1} \binom{k}{k} = \binom{k-1}{r} \quad \text{by (8),} \\ > 0.$$

Hence, no matter how many A_j s occur

$$\begin{aligned} (-)^r \mathbf{E}\left(1 - \prod_{j=1}^n (1 - I_j) - S_1 + S_2 + \cdots + (-)^r S_r\right) \\ = (-)^{2r} \mathbf{E}(S_{r+1} - \cdots + (-)^{t-r+1} S_n) \geq 0, \end{aligned}$$

as required. ■

It can similarly be shown (or deduced from the above) that if

$$t_r = \sum_{i_1 < i_2 < \dots < i_r} \mathbf{P}(A_{i_1} \cup \dots \cup A_{i_r}),$$

then $t_n = \sum_{r=1}^n (-)^{r-1} s_r$, and

$$(-)^r \left(\mathbf{P} \left(\bigcap_1^n A_i \right) - t_1 + t_2 - \dots + (-)^r t_r \right) \geq 0.$$

Often, it is natural and important to consider the sum Z of two variables X and Y . This is itself a random variable, and so we may require the distribution of $Z = X + Y$. This is given by Example 5.1.3, so we have proved that

$$(10) \quad f_Z(z) = \sum_x f_{X,Y}(x, z-x).$$

One special case of this result must be singled out.

(11) **Theorem** *If X and Y are independent discrete random variables, then $Z = X + Y$ has probability mass function*

$$(12) \quad f_Z(z) = \sum_x f_X(x) f_Y(z-x).$$

Proof Substitute Definition 5.2.2 into (10). A summation of this form is called a *convolution*. ■

(13) **Example: Sum of Geometric Random Variables** (a) Let X_1 and X_2 be independent random variables, each having a geometric distribution

$$(14) \quad \mathbf{P}(X_i = k) = (1-p)^k p; \quad k \geq 0.$$

Show that the mass function of $X_1 + X_2$ is $f(z) = (z+1)(1-p)^z p^2; z \geq 0$.

(b) If $(X_i; i > 1)$ are independent random variables, each having the geometric distribution (14), find the mass function of $Z = \sum_1^n X_i$.

Solution (a) Using (12),

$$f(z) = \sum_{k=0}^z (1-p)^k p (1-p)^{z-k} p = \sum_{k=0}^z p^2 (1-p)^z = (z+1) p^2 (1-p)^z.$$

Alternatively, suppose we have a coin that, when tossed, shows a head with probability p . Toss this coin repeatedly until a head first appears, and let T_1 be the number of tails shown up to that point. Continue tossing the coin until the second head appears and let T_2 be the further number of tails shown up to that point. Then, T_1 and T_2 are independent and have the geometric distribution (14), and $T_1 + T_2$ is the number of tails shown before the second head appears. But we know from Example 4.2.8 that

$$\mathbf{P}(T_1 + T_2 = z) = \binom{z+1}{1} p^2 (1-p)^z.$$

Thus, $X_1 + X_2$ also has this mass function.

(b) Extending the second argument above, we see that $\sum_1^n X_i$ has the same distribution as the number of tails shown upto the point where n heads have first appeared in successive tosses of our biased coin. But the probability that z tails have appeared is just

$$(15) \quad \binom{z+n-1}{n-1} p^n (1-p)^z$$

and so this is the mass function of Z . It is negative binomial.

Now that we are in possession of the answer, we can use (12) to verify it by induction. Assume that $\sum_1^n X_i$ has mass function (15). Then, by (12), the mass function of $\sum_1^{n+1} X_i$ is

$$\begin{aligned} \mathbf{P}\left(\sum_{i=1}^{n+1} X_i = z\right) &= \sum_{k=0}^z p(1-p)^k \binom{z+n-k-1}{n-1} (1-p)^{z-k} p^n \\ &= \sum_{k=0}^z p^{n+1} (1-p)^z \binom{z+n-k-1}{n-1} = p^{n+1} (1-p)^z \binom{z+n}{k}, \end{aligned}$$

as required, where we have used the identity

$$(16) \quad \sum_{r=0}^z \binom{n-1+r}{n-1} = \binom{n+z}{n}.$$

Because (15) holds for $n=1$, the result follows by induction. Note that (16) can be derived immediately by equating the coefficient of x^z on both sides of the trivial identity

$$\frac{1}{1+x} \cdot \frac{1}{(1+x)^n} = \frac{1}{(1+x)^{n+1}}. \quad \bullet$$

Example: Sum of Binomial Random Variables Let X and Y be independent random variables with respective mass functions

$$f_X(x) = \binom{m}{x} p^x (1-p)^{m-x}, \quad \text{and} \quad f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}.$$

Show that $X+Y$ has a binomial distribution.

Solution The expeditious method of doing this is to use Example 5.4.1 to write $X = \sum_{i=1}^m I_i$ and $Y = \sum_{i=m+1}^{m+n} I_i$. Hence,

$$X+Y = \sum_{i=1}^{m+n} I_i;$$

this has the $B(m+n, p)$ mass function by Example 5.4.1.

Alternatively, this may be shown by using (5.4.12); this is an exercise for you. \bullet

Turning to expected values, we recall that from Definition 5.2.2, for any random variables $(X_i; i \geq 1)$ with finite expectation, *whether they are independent or not*,

$$(17) \quad \mathbf{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathbf{E}(X_i).$$

If in addition the X_i are independent, then

$$(18) \quad \mathbf{E}\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n \mathbf{E}(X_i).$$

Many inequalities for probabilities and moments are useful when dealing with random vectors. Most of them are beyond our scope at this stage, but we give a few simple examples with some applications.

(19) Basic Inequality If $X \leq Y$ with probability one, then $\mathbf{E}(X) \leq \mathbf{E}(Y)$.

Proof This follows immediately from Theorem 4.3.6. ■

Corollary (a) For $1 < r < s$,

$$(20) \quad \mathbf{E}(|X|^r) \leq (\mathbf{E}(|X|^s)) + 1.$$

(b) For $r \geq 1$,

$$(21) \quad \mathbf{E}(|X + Y|^r) \leq 2^r (\mathbf{E}(|X|^r) + \mathbf{E}(|Y|^r)).$$

Proof (a) If $|x| \leq 1$, then $|x|^r \leq 1$; and if $|x| > 1$ and $r < s$, then $|x|^r \leq |x|^s$. Hence, in any case, when $r \leq s$, $|x|^r \leq |x|^s + 1$. Thus,

$$\mathbf{E}(|X|^r) = \sum_x |x|^r f(x) \leq \sum_x |x|^s f(x) + 1 = \mathbf{E}(|X|^s) + 1.$$

(b) For any real numbers x and y , if $k \leq r$, $|x|^k |y|^{r-k} \leq |x|^r + |y|^r$, because either $(|x|^k / |y|^k) \leq 1$ or $(|y|^{r-k} / |x|^{r-k}) < 1$. Hence,

$$\begin{aligned} |x + y|^r &\leq (|x| + |y|)^r \leq \sum_{k=0}^r \binom{r}{k} |x|^k |y|^{r-k} \\ &\leq \sum_{k=0}^r \binom{r}{k} (|x|^r + |y|^r) = 2^r (|x|^r + |y|^r), \end{aligned}$$

and (3) follows. ■

(22) Corollary These inequalities show that:

(a) If $\mathbf{E}(X^s) < \infty$, then for all $1 \leq r \leq s$, $\mathbf{E}(X^r) < \infty$.

(b) If $\mathbf{E}(X^r)$ and $\mathbf{E}(Y^r)$ are finite then $\mathbf{E}((X + Y)^r) < \infty$. ■

5.5 Dependence: Conditional Expectation

Let X and Y be jointly distributed random variables. We may be given the value of Y , either in fact, or as a supposition. What is the effect on the distribution of X ?

- (1) **Definition** If X and Y have joint probability mass function $f(x, y)$, then given $Y = y$, the random variable X has a conditional probability mass function given by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

for all y such that $f_Y(y) > 0$. ▲

Example Let X and Y be independent geometric random variables each having mass function $f(x) = (1 - \lambda)\lambda^x; x \geq 0, 0 < \lambda < 1$. Let $Z = X + Y$. Show that for $0 \leq x \leq z$, $f_{X|Z}(x|z) = 1/(z + 1)$.

Solution From Example 5.4.13, we know that Z has mass function $f_Z(z) = (z + 1)(1 - \lambda)^2\lambda^z$ and so

$$f_{X|Z}(x|z) = \frac{\mathbf{P}(X = x, Z = z)}{(z + 1)(1 - \lambda)^2\lambda^z} = \frac{(1 - \lambda)^2\lambda^x\lambda^{z-x}}{(1 - \lambda)^2\lambda^z(z + 1)} = (z + 1)^{-1}. \quad \bullet$$

- (2) **Example 5.1.8 Revisited: Cutting for the Deal** Find the conditional mass function of the loser's card conditional on $W = w$; find also $f_{W|V}(w|v)$.

Solution According to Example 5.1.8, $f(v, w) = \frac{1}{78}; 2 \leq v < w \leq 14$, and

$$f_W(w) = \frac{w - 2}{78}; \quad 3 \leq w \leq 14.$$

Hence, using Definition 1,

$$f_{V|W}(v|w) = \frac{1}{w - 2}; \quad 2 \leq v < w.$$

The loser's score is uniformly distributed given W . Likewise,

$$f_{W|V}(w|v) = \frac{1}{78} \bigg/ \frac{14 - v}{78} = \frac{1}{14 - v}; \quad v < w \leq 14,$$

also a uniform distribution. ●

- (3) **Example** Let X and Y be independent. Show that the conditional mass function of X given Y is $f_X(x)$, the marginal mass function.

Solution Because X and Y are independent, $f(x, y) = f_X(x)f_Y(y)$. Hence, applying Definition 1,

$$f_{X|Y}(x|y) = f(x, y)/f_Y(y) = f_X(x). \quad \bullet$$

- (4) **Theorem** $f_{X|Y}(x|y)$ is a probability mass function, which is to say that

- (i) $f_{X|Y}(x|y) \geq 0$ and
- (ii) $\sum_x f_{X|Y}(x|y) = 1$.

Proof Part (i) is trivial. Part (ii) follows immediately from (5.1.5). ■

Recall that two events A and B are said to be conditionally independent given C , if $\mathbf{P}(A \cap B|C) = \mathbf{P}(A|C)\mathbf{P}(B|C)$. Likewise, it is possible for two random variables X and Y to be conditionally independent given Z , if

$$f_{X,Y|Z} = f_{X|Z}f_{Y|Z}.$$

Example: Cutting for the Deal (Example 5.1.8 Revisited) Suppose three players cut for the deal (with ties not allowed, as usual). Let X be the lowest card, Y the highest card and Z the intermediate card. Clearly, X and Y are dependent. However, conditional on $Z = z$, X and Y are independent. The mass function $f_{X|Z}$ is uniform on $\{2, \dots, z-1\}$ and $f_{Y|Z}$ is uniform on $\{z+1, \dots, 14\}$. ●

Being a mass function, $f_{X|Y}$ may have an expectation; it has a special name and importance.

(5) **Definition** The conditional expectation of X , given that $Y = y$ where $f_Y(y) > 0$, is

$$\mathbf{E}(X|\{Y = y\}) = \sum_x xf(x, y)/f_Y(y),$$

when the sum is absolutely convergent. ▲

As y varies over the possible values of Y , this defines a function of Y , denoted by $\mathbf{E}(X|Y)$. Because it is a function of Y , it is itself a random variable, which may have an expectation.

(6) **Theorem** If both sides exist,

(7)
$$\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X).$$

Proof Assuming the sums are absolutely convergent we have, by Theorem 4.3.4,

$$\begin{aligned} \mathbf{E}(\mathbf{E}(X|Y)) &= \sum_y \mathbf{E}(X|\{Y = y\})f_Y(y) = \sum_y \sum_x \frac{xf(x, y)}{f_Y(y)} f_Y(y) \quad \text{by Example 2} \\ &= \sum_x xf_X(x) \quad \text{by (5.1.4)} \\ &= \mathbf{E}(X). \end{aligned}$$

This is an exceptionally important and useful result. Judicious use of Theorem 6 can greatly simplify many calculations; we give some examples.

- (8) **Example: Eggs** A hen lays X eggs where X is Poisson with parameter λ . Each hatches with probability p , independently of the others, yielding Y chicks. Show that $\rho(X, Y) = \sqrt{p}$.

Solution Conditional on $X = k$, the number of chicks is binomial $B(k, p)$, with mean kp . Hence,

$$\mathbf{E}(XY) = \mathbf{E}(\mathbf{E}(XY|X)) = \mathbf{E}(X^2 p) = (\lambda^2 + \lambda)p.$$

Likewise,

$$\begin{aligned}\mathbf{E}(Y^2) &= \mathbf{E}(\mathbf{E}(Y^2|X)) = \mathbf{E}(X_p(1-p) + X^2 p^2) \\ &= \lambda p(1-p) + (\lambda + \lambda^2)p^2 = \lambda p + \lambda^2 p^2.\end{aligned}$$

Hence,

$$\rho(X, Y) = \frac{\mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)}{(\text{var}(X) \text{var}(Y))^{\frac{1}{2}}} = \frac{\lambda^2 p + \lambda p - \lambda \cdot \lambda p}{(\lambda(\lambda p + \lambda^2 p^2 - \lambda^2 p^2))^{\frac{1}{2}}} = \sqrt{p}. \quad \bullet$$

- (9) **Example: Variance of a Random Sum** Let X_1, X_2, \dots be a collection of independent identically distributed random variables, and let Y be an integer valued random variable independent of all the X_i . Let $S_Y = \sum_{i=1}^Y X_i$. Show that

$$\text{var}(S_Y) = \mathbf{E}(X_1)^2 \text{var}(Y) + \mathbf{E}(Y)\text{var}(X_1).$$

Solution By (7),

$$\mathbf{E}(S_Y) = \mathbf{E}(\mathbf{E}(S_Y|Y)) = \mathbf{E}\left(\mathbf{E}\left(\sum_{i=1}^Y X_i | Y\right)\right) = \mathbf{E}(Y\mathbf{E}(X_1)) = \mathbf{E}(Y)\mathbf{E}(X_1).$$

Likewise,

$$\begin{aligned}\mathbf{E}(S_Y^2) &= \mathbf{E}(\mathbf{E}(S_Y^2|Y)) = \mathbf{E}(Y\mathbf{E}(X_1^2) + Y(Y-1)[\mathbf{E}(X_1)]^2) \\ &= \mathbf{E}(Y)(\mathbf{E}(X_1^2) - [\mathbf{E}(X_1)]^2) + \mathbf{E}(Y^2)[\mathbf{E}(X_1)]^2,\end{aligned}$$

and so substituting into $\text{var}(S_Y) = \mathbf{E}(S_Y^2) - (\mathbf{E}(S_Y))^2$, gives the result. ●

- (10) **Example 4.12 Revisited: Gamblers' Ruin** Two gamblers, A and B , have n coins. They divide this hoard by tossing each coin; A gets those that show heads, X say, B gets the rest, totalling $n - X$.

They then play a series of independent fair games; each time A wins he gets a coin from B , each time he loses he gives a coin to B . They stop when one or other has all the coins.

Let D_X be the number of games played. Find $\mathbf{E}(D_X)$, and show that, when the coins are fair, $\rho(X, D_X) = 0$.

Solution Conditional on $X = k$, as in Example 4.12,

$$D_k = \frac{1}{2}D_{k+1} + \frac{1}{2}D_{k-1} + 1$$

with solution $D_k = k(n - k)$. Hence, observing that X is $B(n, p)$ (where p is the chance of a head), we have

$$\mathbf{E}(D_X) = \mathbf{E}(\mathbf{E}(D_X|X)) = \mathbf{E}(X(n - X)) = n(n - 1)p(1 - p).$$

Finally,

$$\text{cov}(X, D_X) = \mathbf{E}(X^2(n - X)) - \mathbf{E}X\mathbf{E}D_X = n(n - 1)p(p - 1)(2p - 1),$$

whence $\rho = 0$, when $p = \frac{1}{2}$

(11) **Example Partition Rule:** Show that if X and Y are jointly distributed, then

$$f_X(x) = \sum_y f_Y(y)f_{X|Y}(x|y). \quad \bullet$$

Solution This is just Theorem 6 in the special case when we take X to be I_x , the indicator of the event $\{X = x\}$. Then, $\mathbf{E}(I_x) = f_X(x)$, and $\mathbf{E}(I_x|Y = y) = f_{X|Y}(x|y)$. The result follows from (7). Alternatively, you can substitute from Definition 1. Essentially this is the Partition Rule applied to discrete random variables. \bullet

Recall that we have already defined $\mathbf{E}(X|B)$ for any event B in Chapter 4. It is convenient occasionally to consider quantities, such as $\mathbf{E}(X|Y; B)$. This is defined to be the expected value of the conditional distribution

$$(12) \quad \mathbf{P}(X = x|Y = y) \cap B = \frac{\mathbf{P}(\{X = x\} \cap \{Y = y\} \cap B)}{\mathbf{P}(\{Y = y\} \cap B)}$$

for any value y of Y such that $\mathbf{P}(\{Y = y\} \cap B) > 0$.

We give some of the more important properties of conditional expectation.

(13) **Theorem** Let a and b be constants, $g(\cdot)$ an arbitrary function, and suppose that X , Y , and Z are jointly distributed. Then (assuming all the expectations exist),

- (i) $\mathbf{E}(a|Y) = a$
- (ii) $\mathbf{E}(aX + bZ|Y) = a\mathbf{E}(X|Y) + b\mathbf{E}(Z|Y)$
- (iii) $\mathbf{E}(X|Y) \geq 0$ if $X \geq 0$
- (iv) $\mathbf{E}(X|Y) = \mathbf{E}(X)$, if X and Y are independent
- (v) $\mathbf{E}(Xg(Y)|Y) = g(Y)\mathbf{E}(X|Y)$
- (vi) $\mathbf{E}(X|Y; g(Y)) = \mathbf{E}(X|Y)$
- (vii) $\mathbf{E}(\mathbf{E}(X|Y; Z)|Y) = \mathbf{E}(X|Y)$.

Property (v) is called the *pull-through property*, for obvious reasons.

Property (vii) is called the *tower property*. It enables us to consider multiple conditioning by taking the random variables in any convenient order.

Proof We prove the odd parts of Theorem 13, the even parts are left as exercises for you.

(i) $f(a, y) = f_Y(y)$, so

$$\mathbf{E}(a|Y) = af_Y(y)/f_Y(y) = a.$$

(iii) If $X \geq 0$, then every term in the sum in Theorem 6 is nonnegative. The result follows.

$$\begin{aligned} \text{(v)} \quad \mathbf{E}(Xg(Y)|Y = y) &= \sum_x xg(y)f(x, y)/f_Y(y) = g(y) \sum_x f(x, y)/f_Y(y) \\ &= g(y)\mathbf{E}(X|Y = y). \end{aligned}$$

(vii) For arbitrary values, $Y = y$ and $Z = z$ of Y and Z , we have $\mathbf{E}(X|Y; Z) = \sum_x xf(x, y, z)/f_{Y,Z}(y, z)$. Hence, by definition,

$$\begin{aligned} \text{(14)} \quad \mathbf{E}(\mathbf{E}(X|Y; Z)|Y) &= \sum_z \mathbf{E}(X|Y; Z)f_{Y,Z}(y, z)/f_Y(y) = \sum_z \sum_x xf(x, y, z)/f_Y(y) \\ &= \sum_x xf(x, y)/f_Y(y) = \mathbf{E}(X|Y). \quad \blacksquare \end{aligned}$$

(15) Example Three children (Aelhyde, Beowulf, and Canute) roll a die in the order A, B, C, A, \dots , etc. until one of them rolls a six and wins. Find the expected number of rolls, given that Canute wins.

Solution We use a form of the tower property, (14). Let C be the event that Canute wins, let X be the duration of the game, and let Y denote the first roll to show a six in the first three rolls, with $Y = 0$ if there is no six. Then,

$$\text{(16)} \quad \mathbf{E}(X|C) = \mathbf{E}(\mathbf{E}(X|Y; C)|C).$$

Now if $Y = 0$, then with the fourth roll the game stands just as it did initially, except that three rolls have been made. So

$$\mathbf{E}(X|Y = 0; C) = 3 + \mathbf{E}(X|C).$$

Obviously, Y is otherwise 3, and

$$\mathbf{E}(X|Y = 3; C) = 3.$$

Therefore, substituting in (16), we have

$$\mathbf{E}(X|C) = 3 + \mathbf{E}(X|C)\mathbf{P}(Y = 0|C)$$

and so

$$\mathbf{E}(X|C) = \frac{3}{1 - \left(\frac{5}{6}\right)^3} = \frac{648}{91}.$$

Of course, there are other ways of doing this. ●

Finally, we remark that conditional expectation arises in another way.

(17) Theorem Let $h(Y)$ be any function of Y such that $\mathbf{E}(h(Y)^2) < \infty$. Then,

$$\text{(18)} \quad \mathbf{E}((X - h(Y))^2) \geq \mathbf{E}((X - \mathbf{E}(X|Y))^2).$$

Further, if $h(Y)$ is any function of Y such that

$$(19) \quad \mathbf{E}((X - h(Y))^2) = \mathbf{E}((X - \mathbf{E}(X|Y))^2),$$

then

$$\mathbf{E}((h(Y) - \mathbf{E}(X|Y))^2) = 0.$$

Proof

$$(20) \quad \begin{aligned} \mathbf{E}((X - h(Y))^2) &= \mathbf{E}((X - \mathbf{E}(X|Y) + \mathbf{E}(X|Y) - h(Y))^2) \\ &= \mathbf{E}((X - \mathbf{E}(X|Y))^2) + \mathbf{E}((\mathbf{E}(X|Y) - h(Y))^2) \\ &\quad + 2\mathbf{E}((X - \mathbf{E}(X|Y))(\mathbf{E}(X|Y) - h(Y))). \end{aligned}$$

However, by (7), we can write

$$\begin{aligned} \mathbf{E}((X - \mathbf{E}(X|Y))(\mathbf{E}(X|Y) - h(Y))) &= \mathbf{E}(\mathbf{E}((X - \mathbf{E}(X|Y))(\mathbf{E}(X|Y) - h(Y))|Y)) \\ &= \mathbf{E}(\mathbf{E}(X|Y) - h(Y))\mathbf{E}(\mathbf{E}(X|Y)|Y) \\ &\hspace{15em} \text{by Theorem 13(v)} \\ &= 0, \end{aligned}$$

because $\mathbf{E}((X - \mathbf{E}(X|Y))|Y) = 0$. The result (18) follows because $\mathbf{E}((\mathbf{E}(X|Y) - h(Y))^2) \geq 0$. Finally, if (19) holds then from (20), we have

$$\mathbf{E}((\mathbf{E}(X|Y) - h(Y))^2) = 0,$$

as required to complete the proof of Theorem 17. ■

Recall that if $\mathbf{E}(X^2) = 0$, then $\mathbf{P}(X = 0) = 1$; hence, if $\mathbf{E}((X - Y)^2) = 0$, then $X = Y$ with probability one. This suggests that the smaller $\mathbf{E}((X - Y)^2)$ is, then the “closer” X is to Y , in some sense. The point of the theorem is then that among all functions of Y , $\mathbf{E}(X|Y)$ is the one which is “closest” to X .

It is thus possible, and in later work desirable, to define $\mathbf{E}(X|Y)$ by this property. However, to explore all these ideas would take us too far afield.

5.6 Simple Random Walk

The ideas above now enable us to consider an exceptionally famous and entertaining collection of random variables.

- (1) **Definition** *Let $(X_i; i \geq 1)$ be a collection of independent identically distributed random variables with mass function*

$$\mathbf{P}(X_1 = 1) = p; \mathbf{P}(X_1 = -1) = q = 1 - p.$$

Then the collection $(S_n; n \geq 0)$, where

$$(2) \quad S_n = S_0 + \sum_{i=1}^n X_i,$$

is called a simple random walk. If $p = q$, it is called a symmetric simple random walk.



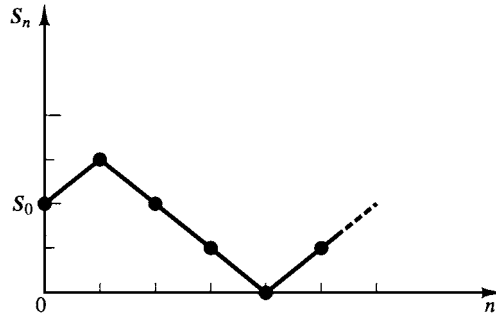


Figure 5.1 A path (or realization) of a simple random walk. The points represent successive positions (or values) of the walk; these are joined by steps of the walk. In this case, $S_0 = 2$, $S_1 = 3$, $S_2 = 2$, and so on. The walk visits the origin at the fourth step.

The nomenclature follows from the visualization of S_n as representing the position of a particle that is initially at S_0 , and then takes a series of independent unit steps; each step being positive with probability p or negative with probability q .

It is conventional to display the walk in Cartesian coordinates as the sequence of points (n, S_n) for $n \geq 0$. Any particular such sequence is called a path of the random walk; see Figure 5.1 for an illustration.

Much of the effort of the first probabilists (Fermat, Pascal, Bernoulli, de Moivre, Laplace) was devoted to discovering the properties of the simple random walk, and more general random walks are still being investigated by modern probabilists.

It is easy to see that the celebrated gambler's ruin problem of Example 2.11 is just a simple random walk in which S_0 is interpreted as the initial capital of the gambler, and the walk stops on the first occasion D when either $S_n = 0$ (the gambler is ruined) or $S_n = K$ (his opponent is ruined). That is, the random variable

$$(3) \quad D = \min\{n : \{S_n = 0\} \cup \{S_n = K\}\}$$

is the duration of the game. In the context of random walks, D is called the first passage time of the walk to $\{0, K\}$.

The first thing to find is the mass function of S_n (when it is not stopped).

(4) Theorem

$$\mathbf{P}(S_n - S_0 = k) = \binom{n}{\frac{1}{2}(n+k)} p^{\frac{1}{2}(n+k)} q^{\frac{1}{2}(n-k)}.$$

Proof Consider a path of the walk from $(0, S_0)$ to (n, S_n) with r positive steps and s negative steps. If $S_n - S_0 = k$, then $r - s = k$ and $r + s = n$. Hence, $r = \frac{1}{2}(n+k)$ and $s = \frac{1}{2}(n-k)$. There are $\binom{n}{r}$ such paths, and each has the same probability, namely, $p^r q^s$. Hence,

$$\mathbf{P}(S_n - S_0 = k) = \binom{n}{r} p^r q^s,$$

which is (4).

Alternatively, we can simply observe that $\frac{1}{2}(S_n - S_0 + n)$ is a binomial random variable with parameters n and p , and (4) follows. ■

As suggested by the gambler's ruin problem, we are interested in the first passage times of random walks.

- (5) **Definition** Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = i$. Then, the first passage time from i to k is the random variable

$$T_{ik} = \min\{n > 0; S_n = k\}. \quad \blacktriangle$$

When $i = k$, the random variable T_{kk} is called the *recurrence time* of k . We often denote T_{kk} by T_k . One obvious but important property of a first passage time T_{ik} is that steps after the first passage to k are independent of those before. It follows that we can write, for example,

$$(6) \quad T_{02} = T_{01} + T_{12},$$

where T_{01} and T_{12} are independent. Furthermore, T_{12} and T_{01} have the same distribution because the X_i are identically distributed. These simple remarks are of great importance in examining the properties of the walk.

Our first result is striking.

- (7) **Theorem** If $p \geq q$, then T_{01} is certainly finite. If $p < q$, then T_{01} is finite with probability p/q .

Proof Let $\mathbf{P}(T_{jk} < \infty) = r_{jk}$. Now let us condition on the first step of the walk from $S_0 = 0$, giving

$$(8) \quad r_{01} = \mathbf{P}(T_{01} < \infty) = p\mathbf{P}(T_{01} < \infty | X_1 = 1) + q\mathbf{P}(T_{01} < \infty | X_1 = -1).$$

On the one hand, if $X_1 = 1$, then $T_{01} = 1$. On the other hand, if $X_1 = -1$, then the walk has to go from -1 to 0 and then from 0 to $+1$ in a finite number of steps for $T_{01} < \infty$. Hence, as $r_{01} = r_{-10}$,

$$(9) \quad r_{01} = p + qr_{01}^2.$$

This has two roots $r_{01} = 1$ and $r_{01} = p/q$. It follows that if $p \geq q$, then the only root that is a probability is $r_{01} = 1$, as required.

If $p < q$, then it certainly seems plausible that $r_{01} = p/q$, but a little work is needed to prove it. This can be done in several ways. We choose to use the following interesting fact; that is, for all i ,

$$(10) \quad \mathbf{E} \left(\left(\frac{q}{p} \right)^{X_i} \right) = p \left(\frac{q}{p} \right)^{+1} + q \left(\frac{q}{p} \right)^{-1} = 1.$$

Hence, if $S_0 = 0$,

$$(11) \quad \mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \right) = \prod_{i=1}^n \mathbf{E} \left(\left(\frac{q}{p} \right)^{X_i} \right) = 1.$$

Now denote T_{01} by T , and suppose that

$$(12) \quad p < q \quad \text{and} \quad r_{01} = 1.$$

Then conditioning on whether $T \leq n$ or $T > n$ gives

$$(13) \quad \begin{aligned} 1 &= \mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \right) \\ &= \mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \mid T \leq n \right) \mathbf{P}(T \leq n) + \mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \mid T > n \right) \mathbf{P}(T > n). \end{aligned}$$

Now if $T \leq n$, then $S_T = 1$ and $S_n = 1 + X_{T+1} + \cdots + X_n$. Hence,

$$\mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \mid T \leq n \right) = \frac{q}{p} \mathbf{E} \left(\left(\frac{q}{p} \right)^{X_{T+1} + \cdots + X_n} \right) = \frac{q}{p} \quad \text{by (10).}$$

Furthermore, if $T > n$, then $S_n \leq 0$, and so

$$\mathbf{E} \left(\left(\frac{q}{p} \right)^{S_n} \mid T > n \right) \leq 1.$$

Hence, allowing $n \rightarrow \infty$ in (13) gives $1 = q/p + 0$. But this is impossible when $p < q$, so (12) must be impossible. Hence, when $p < q$, we must have $r_{01} = p/q$. ■

In the case when $p > q$, so that T_{01} is finite, it is natural to ask what is $\mathbf{E}(T_{01})$? If we knew that $\mathbf{E}(T_{01}) < \infty$, then we could write

$$\begin{aligned} \mathbf{E}(T_{01}) &= \mathbf{E}(\mathbf{E}(T_{01} | X_1)) = p\mathbf{E}(T_{01} | X_1 = 1) + q\mathbf{E}(T_{01} | X_1 = -1) \\ &= p + q(1 + \mathbf{E}(T_{-1,1})) = 1 + 2q\mathbf{E}(T_{01}) \quad \text{by (6)}. \end{aligned}$$

Hence,

$$(14) \quad \mathbf{E}(T_{01}) = \frac{1}{p - q}; \quad p > q,$$

as required.

It is not too difficult to show that $\mathbf{E}(T_{01}) < \infty$, as we now demonstrate.

(15) Theorem $\mathbf{E}(T_{01}) < \infty$ when $p > q$.

Proof

$$\begin{aligned} \mathbf{P}(T_{01} > n) &= \mathbf{P}(S_i \leq 0 \text{ for } 0 \leq i \leq n) \\ &\leq \mathbf{P}(S_n \leq 0) = \mathbf{P}\left(\frac{1}{2}(S_n + n) \leq \frac{n}{2}\right) \end{aligned}$$

$$\begin{aligned}
&= \mathbf{P} \left(\left(\frac{q}{p} \right)^{\frac{1}{2}(S_n+n)} \geq \left(\frac{q}{p} \right)^{n/2} \right) \text{ since } p > q, \\
&\leq \mathbf{E} \left(\left(\frac{q}{p} \right)^{\frac{1}{2}(S_n+n)} \right) \left(\frac{q}{p} \right)^{-n/2}
\end{aligned}$$

by the basic inequality, Theorem 4.6.1.

Now we recall the observation in Theorem 4 that $\frac{1}{2}(S_n + n)$ has the $B(n, p)$ mass function. Hence,

$$(16) \quad \mathbf{E} \left(\left(\frac{q}{p} \right)^{\frac{1}{2}(S_n+n)} \right) = \left(q + p \cdot \frac{q}{p} \right)^n = (2q)^n.$$

Therefore, finally,

$$\mathbf{E}(T_{01}) = \sum_{n=0}^{\infty} \mathbf{P}(T_{01} > n) \leq \sum_{n=0}^{\infty} (2q)^n \left(\frac{p}{q} \right)^{n/2} = \frac{1}{1 - 2(pq)^{\frac{1}{2}}} \quad \text{since } pq < \frac{1}{4}$$

This establishes (14), as required. \blacksquare

There are one or two gaps in the above; for example, we do not yet know either the mass function of T_{01} or $\mathbf{E}(T_{01})$ in the case $p = q = \frac{1}{2}$. Both of these can be filled by the following beautiful theorem.

(17) Hitting Time Theorem Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = 0$. Let T_{0b} be the first passage time from 0 to $b > 0$, with mass function $f_{0b}(n)$. Then

$$f_{0b}(n) = \mathbf{P}(T_{0b} = n) = \frac{b}{n} \mathbf{P}(S_n = b) = \frac{b}{n} \binom{n}{\frac{1}{2}(n+b)} p^{\frac{1}{2}(n+b)} q^{\frac{1}{2}(n-b)}.$$

The proof of (17) relies on the following lemma, which is of considerable interest in its own right. First, we observe that the number of paths of the walk from $(0, 0)$ to $(n-1, b+1)$ is denoted by $N_{n-1}(0, b+1)$, and we have

$$(18) \quad N_{n-1}(0, b+1) = \binom{n-1}{\frac{1}{2}(n-b)-1}.$$

(19) Lemma: The Reflection Principle Let $N_{n-1}^b(0, b-1)$ be the number of paths from $(0, 0)$ to $(n-1, b-1)$ that pass through b at least once. Then

$$(20) \quad N_{n-1}^b(0, b-1) = N_{n-1}(0, b+1). \quad \bullet$$

Proof Let π be a path that visits b on its journey from $(0, 0)$ to $(n-1, b-1)$. Let L be the occasion of its last visit. Now reflect that part of the walk after L in the line $y = b$. This yields a path π' from $(0, 0)$ to $(n-1, b+1)$. Conversely, for any path from $(0, 0)$ to $(n-1, b+1)$, we may reflect the segment in $y = b$ after its last visit

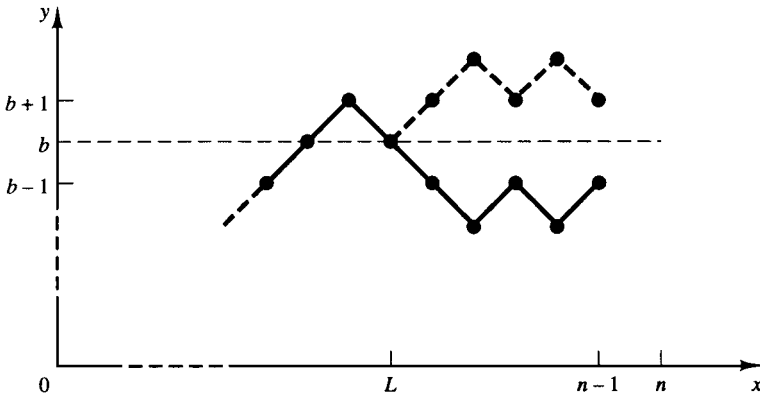


Figure 5.2 The solid line is the path of the walk: the dashed line is the reflection in $y = b$ of that part of the walk after its last visit to b before n , at time L .

to b , to give a path π from $(0, 0)$ to $(n - 1, b - 1)$. These two sets are thus in one-one correspondence, and (20) follows. Figure 5.2 illustrates the reflection. ■

Proof of (17): Hitting Time Theorem If $T_{0b} = n$, then we must have $X_n = +1$ and $S_{n-1} = b - 1$. Now there are $N_{n-1}(0, b - 1)$ paths from $(0, 0)$ to $(n - 1, b - 1)$ of which $N_{n-1}^b(0, b - 1)$ visit b on route. Each such path has probability $p^{\frac{1}{2}(n+b)-1}q^{\frac{1}{2}(n-b)}$. Hence, using the reflection principle,

$$\begin{aligned} \mathbf{P}(T_{0b} = n) &= p(N_{n-1}(0, b - 1) - N_{n-1}(0, b + 1))p^{\frac{1}{2}(n+b)-1}q^{\frac{1}{2}(n-b)} \\ &= \left(\binom{n-1}{\frac{1}{2}(n+b)-1} - \binom{n-1}{\frac{1}{2}(n+b)} \right) p^{\frac{1}{2}(n+b)}q^{\frac{1}{2}(n-b)} \\ &= \frac{b}{n} \binom{n}{\frac{1}{2}(n+b)} p^{\frac{1}{2}(n+b)}q^{\frac{1}{2}(n-b)} \\ &= \frac{b}{n} \mathbf{P}(S_n = b), \quad \text{by (4)}. \end{aligned}$$

Because a similar argument works for negative values of b , we have

$$(21) \quad \mathbf{P}(T_{0b} = n) = \frac{|b|}{n} \mathbf{P}(S_n = b)$$

and $\mathbf{E}(T_{0b}) = \sum_{n=1}^{\infty} |b| \mathbf{P}(S_n = b)$.

(22) Example: Symmetric Random Walk When $p = q$, the simple random walk is said to be symmetric. In this case,

$$\begin{aligned} \mathbf{E}(T_{01}) &= \sum_{m=0}^{\infty} \mathbf{P}(S_{2m+1} = 1) = \sum_{m=0}^{\infty} \binom{2m+1}{m+1} 2^{-(2m+1)} \\ &\geq \sum_{m=0}^{\infty} \frac{2m(2m-2)\dots 2 \cdot 1}{(m+1)(m)\dots 1} 2^{-m} = \sum_{m=0}^{\infty} \frac{1}{m+1} = \infty. \end{aligned}$$

Hence, the symmetric random walk has the interesting property that $\mathbf{P}(T_{01} < \infty) = 1$, but

$$(23) \quad \mathbf{E}(T_{01}) = \infty. \quad \bullet$$

(24) **Example: Conditioned Random Walk** Now, of course, when $p < q$ we know that $\mathbf{P}(T_{01} < \infty) = p/q < 1$, so T_{01} has no expectation. But consider S_n conditional on the event that $T_{01} < \infty$. In this case, by conditional probability,

$$\mathbf{P}(X_1 = +1 | T_{01} < \infty) = \frac{\mathbf{P}(T_{01} < \infty | X_1 = 1)\mathbf{P}(X_1 = 1)}{\mathbf{P}(T_{01} < \infty)} = p / \left(\frac{p}{q}\right) = q.$$

Likewise, $\mathbf{P}(X_1 = -1 | T_{01} < \infty) = p$. Hence, if we knew that $\mathbf{E}(T_{01} | T_{01} < \infty)$ were finite, then by conditioning on the first step,

$$\begin{aligned} \mathbf{E}(T_{01} | T_{01} < \infty) &= q + p\mathbf{E}(T_{01} | X_1 = -1; T_{01} < \infty) \\ &= 1 + p2\mathbf{E}(T_{01} | T_{01} < \infty), \quad \text{by (6)}. \end{aligned}$$

Hence, when $q > p$,

$$(25) \quad \mathbf{E}(T_{01} | T_{01} < \infty) = \frac{1}{q - p}.$$

It is straightforward to use Theorem 17 to establish that $\mathbf{E}(T_{01} | T_{01} < \infty) < \infty$ (an exercise for you) and (25) is proved. Together with (14) and Example 22, this shows that for any value of p and $b > 0$,

$$(26) \quad \mathbf{E}(T_{0b} | T_{0b} < \infty) = \frac{b}{|p - q|}. \quad \bullet$$

We may also consider recurrence times.

(27) **Example** Let T_0 be the recurrence time of 0. Show that $\mathbf{P}(T_0 < \infty) = 1 - |p - q|$ and

$$\mathbf{E}(T_0 | T_0 < \infty) = 1 + \frac{1}{|p - q|}.$$

Solution Just consider T_0 conditional on the outcome of the first step, and then use what we know about T_{01} and T_{10} . You fill in the details. ●

Finally, we prove a famous result, the so-called

(28) **Ballot Theorem** Let $S_n = \sum_1^n X_i$ be a simple random walk with $S_0 = 0$. Then

$$\mathbf{P}\left(\prod_1^{2n-1} S_i \neq 0 | S_{2n} = 2r\right) = \frac{r}{n}.$$

Proof We count paths as we did in the hitting time theorem. What is the number $N_{2n-1}^0(1, 2r)$ of paths from $(1, 1)$ to $(2n, 2r)$ that visit the origin? We can reflect the

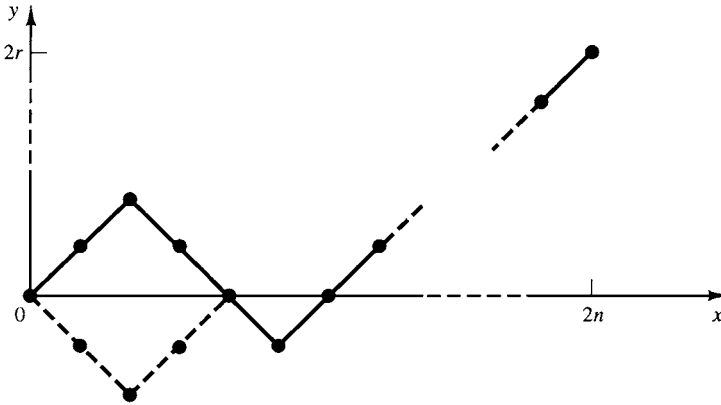


Figure 5.3 The ballot theorem The solid line is a path of the walk; the dashed line is the reflection in the x -axis of that part of the walk before its first visit to zero.

walk before its first zero in the x -axis, see Figure 5.3, and this shows that $N_{2n-1}^0(1, 2r) = N_{2n-1}(-1, 2r)$. Because all $N_{2n}(0, 2r)$ paths from $(0, 0)$ to $(2n, 2r)$ are equally likely, it follows that the required probability is

$$\begin{aligned} \frac{N_{2n-1}(1, 2r) - N_{2n-1}^0(1, 2r)}{N_{2n}(0, 2r)} &= \frac{N_{2n-1}(1, 2r) - N_{2n-1}(-1, 2r)}{N_{2n}(0, 2r)} \\ &= \frac{\binom{2n-1}{n+r-1} - \binom{2n-1}{n+r}}{\binom{2n}{n+r}} = \frac{2r}{2n}. \quad \blacksquare \end{aligned}$$

The following application explains the name of the theorem.

Example: Ballot In an election, candidate A secures a votes and candidate B secures b votes. What is the probability that A is ahead throughout the count? By the above argument, this probability is $(a - b)/(a + b)$ when $a > b$. ●

5.7 Martingales

In this section, we consider a remarkably useful class of random processes called martingales. They arise naturally as general models for fair games, but turn up in all kinds of unexpected places. In particular, they are used extensively in modern financial mathematics, but it is beyond our scope to explore this area in great detail. We begin with this:

- (1) **Definition** A collection $(S_n; n \geq 0)$ of random variables is a martingale if, for all n ,
- (a) $\mathbf{E}|S_n| < \infty$.
 - (b) $\mathbf{E}(S_{n+1} | S_0, S_1, \dots, S_n) = S_n$. ▲

This definition clearly shows the interpretation as a fair game; if S_n is a gambler's fortune after the n th play, then (b) asserts that the expectation of this fortune after the next play—taking into account all previous fluctuations in his fortune—is simply equal to S_n . Briefly, conditional on the past, future expectations equal the current value. Note that this section will rely heavily on the properties of conditional expectations summarized in 5.5.13; keep them well in mind.

Martingales get their name from a particularly well-known gambling strategy that we discussed above. We recall example 4.15 in the special case when $p = \frac{1}{2}$.

- (2) **Example: The Martingale** You bet \$1 at evens (calling heads on the flip of a fair coin, say); if you win you quit. If you lose, you bet \$2 at evens, and so on. That is, you double the stake at each loss, and quit at the first win. Let S_n denote your fortune after the n th bet, and let X_n denote the outcome of the n th flip of the coin; thus,

$$X_n = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Because $\mathbf{E}X_{n+1} = 0$, it follows immediately that $\mathbf{E}(S_{n+1}|S_0, \dots, S_n) = S_n$. Also,

$$|S_n| \leq 1 + 2 + \dots + 2^n \leq 2^{n+1},$$

so $\mathbf{E}|S_n| < \infty$ and S_n is a martingale. If the game stops at the n th flip, your fortune is $S_n = -1 - 2 - 4 - \dots - 2^{n-1} + 2^n = 1$, so you always win.

However, recall that this game is “fair” only in a mathematical sense, and the strategy is not as good as it may look! It has the serious drawback that the expected size of the winning bet is infinite. To see this, note that the game ends on the n th play with probability 2^{-n} . The stake on this play is \$ 2^{n-1} . So the expected stake is \$ $\sum_{r=0}^{\infty} 2^{-r} \cdot 2^{r-1}$, which is infinite. You are well-advised not to gamble, and above all avoid the martingale if you do. ●

We give a few simple examples of natural martingales.

- (3) **Example** Let $(X_n; n \geq 0)$ be independent.

(a) If $\mathbf{E}X_n = 0$, then $S_n = \sum_{r=0}^n X_r$ defines a martingale because

$$\mathbf{E}(S_{n+1}|S_0, \dots, S_n) = \mathbf{E}(S_n + X_{n+1}|S_n) = S_n.$$

(b) If $\mathbf{E}X_n = 1$, then $S_n = \prod_{r=0}^n X_r$ defines a martingale because

$$\mathbf{E}(S_{n+1}|S_0, \dots, S_n) = \mathbf{E}(S_n X_{n+1}|S_n) = S_n. \quad \bullet$$

The properties of conditional expectation extend in natural and obvious ways to larger collections of random variables; we single out, in particular, the tower property 5.5.13 vii in the presence of random vectors \mathbf{Y} and \mathbf{Z} :

- (4)
$$\mathbf{E}(X|\mathbf{Y}) = \mathbf{E}(\mathbf{E}(X|\mathbf{Y}, \mathbf{Z})|\mathbf{Y}).$$

This leads to an important class of martingales.

- (5) **Example: Doob Martingale** Let X, X_0, X_1, \dots be any collection of jointly distributed random variables with $\mathbf{E}|X| < \infty$. Define

$$M_n = \mathbf{E}(X|X_0, X_1, \dots, X_n).$$

Then by Jensen's inequality, see Theorem 4.6.14 and Problem 5.42,

$$\begin{aligned} \mathbf{E}|M_n| &= \mathbf{E}[|\mathbf{E}(X|X_0, \dots, X_n)|] \\ &\leq \mathbf{E}[\mathbf{E}(|X||X_0, \dots, X_n)] = \mathbf{E}|X| < \infty, \end{aligned}$$

and

$$\begin{aligned} \mathbf{E}(M_{n+1}|X_0, \dots, X_n) &= \mathbf{E}(\mathbf{E}(X|X_0, \dots, X_{n+1})|X_0, \dots, X_n) \\ &= \mathbf{E}(X|X_0, \dots, X_n) \quad \text{by (4)} \\ &= M_n. \end{aligned}$$

Hence, M_n is a martingale. ●

As one example of this, consider the following example.

- (6) **Example: An Options Martingale** Suppose that $(X_n; n \geq 0)$ represent the price of some stock on successive trading days, $n \geq 0$. Naturally, $\mathbf{E}|X_n| < \infty$. Suppose you own the right (but not the obligation) to purchase this stock at a fixed price K at an exercise date T . Then your option at that date is worth X , say, where

$$X = \max\{(X_T - K), 0\} = (X_T - K)^+,$$

because if $X_T < K$, then your option is worthless; you could buy the stock anyway for the actual price less than K . At any time $n < T$, the expected value of your option, in the knowledge of the stock prices up to then, is

$$M_n = \mathbf{E}(X|X_0, \dots, X_n).$$

By the previous example, M_n is a martingale. [Note well that M_n is not the fair price for this option!] ●

We will see many more examples of martingales later, but for the moment we turn aside to observe that the key property of all realistic fair games is that they have to stop. For example, the gambler is bankrupt or decides to quit while ahead, or the casino imposes a house limit. There are many other real life actions that have this central property that their nature is fixed, but their timing is optional. In all such cases, the action can only be taken in light of your knowledge *up to the time of execution*. Nobody can follow a rule that says "stop just before you have a big loss." This is unreasonable.

We therefore make a useful definition of such reasonable times of action, called *stopping times*, as follows.

- (7) **Definition** A random variable T taking values in $\{0, 1, 2, \dots\}$ is called a stopping time with respect to $\{S_0, S_1, \dots\}$, if the event $\{T = n\}$ may depend only on $\{S_0, S_1, \dots, S_n\}$, and is independent of S_{n+k} for all $k \geq 1$. It may be that $\mathbf{P}(T < \infty) = 1$, in which case T is said to be almost surely finite. ▲

- (8) **Example: The Martingale 4.15 Again** The wheel is spun repeatedly and, in this case, T is the time when it first yields red. Let I_k be the indicator of the event that the k th spin yields red, so that by construction

$$\{T = n\} = \{I_k = 0, 1 \leq k \leq n-1, I_n = 1\}.$$

Note that $\{T = n\}$ does not depend on any I_{n+k} , so T is a stopping time for the martingale. In this case, $\mathbf{P}(T < \infty) = \lim_{k \rightarrow \infty} \{1 - \mathbf{P}(T > k)\} = 1 - \lim_{k \rightarrow \infty} (1-p)^k = 1$, so T is almost surely finite. ●

- (9) **Example: First Passage Times** If $\{S_n; n \geq 0\}$ is a random walk, with $S_0 = 0$ (say) and $T = \min\{n : S_n \geq b\}$, then T is easily seen to be a stopping time. ●

Stopping times are at least as important to the theory of martingales as they are in real life. The main reason for this is the fact that a martingale $\{X_n; n \geq 0\}$ that is stopped at a random time T is still a martingale, provided that T is a stopping time for $\{X_n; n \geq 0\}$. That is to say, formally:

- (10) **Theorem** Let T be a stopping time for the martingale $\{X_n; n \geq 0\}$, and let

$$(11) \quad Z_n = X_{T \wedge n} = \begin{cases} X_n, & n \leq T \\ X_T, & n > T \end{cases}$$

Then Z_n is a martingale, and

$$(12) \quad \mathbf{E}Z_n = \mathbf{E}X_0.$$

Proof We can rewrite Z_n using indicators as

$$Z_n = \sum_{r=0}^{n-1} X_r I\{T = r\} + X_n I\{T \geq n\}.$$

Hence, $\mathbf{E}|Z_n| \leq \sum_0^n \mathbf{E}|X_r| < \infty$. Also, using indicators again,

$$Z_{n+1} = Z_n + (X_{n+1} - X_n)I\{T > n\}.$$

To see this, note that if $I\{T > n\} = 1$, then $Z_{n+1} = X_{n+1}$, and if $I\{T \leq n\} = 1$, then $Z_{n+1} = Z_n$, both being consonant with the definition of Z_n .

Next we note that, because $\{T > n\}$ is independent of X_{n+k} for all k , and is a function of X, \dots, X_n ,

$$(13) \quad \begin{aligned} \mathbf{E}[(X_{n+1} - X_n)I(T > n)|X_0, \dots, X_n] \\ = I(T > n)\mathbf{E}(X_{n+1} - X_n|X_0, \dots, X_n) = 0, \end{aligned}$$

using the pull through property (5.5.13)(v). Hence,

$$\mathbf{E}(Z_{n+1}|Z_0, \dots, Z_n) = \mathbf{E}(Z_{n+1}|X_0, \dots, X_n) = Z_n,$$

and the result follows. ■

Now, if T is almost surely finite, it is true with probability 1 that $Z_n \rightarrow X_T$ as $n \rightarrow \infty$, from (11). It is natural to ask if also, as $n \rightarrow \infty$,

$$\mathbf{E}Z_n \rightarrow \mathbf{E}X_T,$$

which would entail, using (12), the remarkable result that

$$\mathbf{E}X_T = \mathbf{E}X_0.$$

It turns out that this is true, under some extra conditions. Here are some popular cases:

(14) **Theorem: Optional stopping** *Let X_n be a martingale and T a stopping time for $(X_n; n \geq 0)$. Then $\mathbf{E}X_T = \mathbf{E}X_0$, if any of the following hold for some positive finite constant K .*

- (a) T is bounded (i.e., $T \leq K < \infty$).
- (b) $|X_n| \leq K$ for all n , and $\mathbf{P}(T < \infty) = 1$.
- (c) $\mathbf{E}|X_{n+1} - X_n||X_0, \dots, X_n| \leq K$ for $n < T$, and $\mathbf{E}T < \infty$.

Proof We prove (a) and (b) here, postponing the proof of (c) to Theorem 5.9.9. First, recall that we showed $\mathbf{E}X_{T \wedge n} = \mathbf{E}X_0$ in Theorem (10). So if we take $n = K$, this proves (a) is sufficient.

To show that (b) is sufficient, note that

$$\begin{aligned} |\mathbf{E}X_0 - \mathbf{E}X_T| &= |\mathbf{E}X_{T \wedge n} - \mathbf{E}X_T| \\ &\leq 2K\mathbf{P}(T > n), \text{ because } |X_n| < K, \\ &\rightarrow 0 \text{ as } n \rightarrow \infty, \text{ because } \mathbf{P}(T < \infty) = 1. \end{aligned}$$

Hence, $|\mathbf{E}X_0 - \mathbf{E}X_T| = 0$, which yields the result. ■

This simple-looking result is remarkably useful and powerful; we give many illustrations of this in later worked examples. Here is one to begin with.

- (15) **Example: Wald's Equation** Let $(X_n; n \geq 1)$ be independent, having the common mean μ , and set

$$Y_n = Y_0 + \sum_{r=1}^n X_r - n\mu.$$

It is easy to see that Y_n is a martingale (when $\mathbf{E}|Y_0| < \infty$) because

$$\mathbf{E}(Y_{n+1}|Y_0, \dots, Y_n) = Y_n + \mathbf{E}(X_{n+1} - \mu).$$

Furthermore,

$$\mathbf{E}|Y_{n+1} - Y_n| = \mathbf{E}|X_{n+1} - \mu| \leq \mathbf{E}|X_1| + |\mu| < \infty.$$

Hence, when $Y_0 = 0$, and if T is any stopping time for $(Y_n, n \geq 0)$ such that $\mathbf{E}T < \infty$, part (c) of the optional stopping theorem yields

$$\mathbf{E}Y_T = \mathbf{E}\left(\sum_{r=1}^T X_r - T\mu\right) = \mathbf{E}Y_0 = 0.$$

That is to say,

$$(16) \quad \mathbf{E}\sum_{r=1}^T X_r = \mu\mathbf{E}T.$$

Of course, this would be trivial when T is independent of the X_n . It is remarkable that it remains true when T depends on the sequence X_n . ●

We conclude this section by noting that there are several other kinds of interesting and important martingales. Recall that a martingale is intuitively your fortune at the n th play of some fair game, and the martingale property implies that, for the martingale X_n , $\mathbf{E}(X_{n+1} - X_n|X_0, \dots, X_n) = 0$. It is equally easy to show that this implies the martingale property. The extension to unfair games is natural:

- (17) **Definition** Suppose that the sequence X_n satisfies $\mathbf{E}|X_n| < \infty$. Then

- (a) X_n is a supermartingale if $\mathbf{E}(X_{n+1} - X_n|X_0, \dots, X_n) \leq 0$.
 (b) X_n is a submartingale if $\mathbf{E}(X_{n+1} - X_n|X_0, \dots, X_n) \geq 0$.

These correspond to unfavourable and favourable games, respectively. ▲

A different type of martingale is motivated by looking at the average of a random walk.

- (18) **Example: Backward Martingale** Let $S_n = \sum_1^n X_r$, where the X_r are independent and identically distributed with common mean $\mu < \infty$. Define

$$M_1 = \frac{S_n}{n}, M_2 = \frac{S_{n-1}}{n-1}, \dots, M_n = \frac{X_1}{1},$$

which is to say that $M_m = S_{n-m+1}/(n-m+1)$. Then obviously, $\mathbf{E}|M_n| < \infty$. Also by symmetry, for $1 \leq r \leq m$,

$$\mathbf{E}(X_r|S_m) = S_m/m.$$

Hence, also by symmetry,

$$\begin{aligned} \mathbf{E}(M_{m+1}|M_1, \dots, M_m) &= \mathbf{E}\left(\frac{X_1 + \dots + X_{n-m}}{n-m} | S_{n-m+1}\right) \\ &= \frac{S_{n-m+1}}{n-m+1} = M_m. \end{aligned}$$

Therefore, M_m , $1 \leq m \leq n$ is a martingale, and so the sequence

$$Y_r = \frac{S_r}{r}, \quad 1 \leq r \leq n$$

is called a backward martingale for obvious reasons. ●

It is often convenient to make a slightly more general definition of a martingale.

(19) Definition A sequence $(S_n; n \geq 0)$ is a martingale with respect to the sequence $(X_n; n \geq 0)$ if $\mathbf{E}|S_n| < \infty$ and

$$\mathbf{E}(S_{n+1}|X_0, \dots, X_n) = S_n,$$

for all n . ▲

In fact, we often omit any reference to the underlying sequence $(X_n; n \geq 0)$, but simply say that $(S_n; n \geq 0)$ is a martingale.

Finally, we note that there are appropriate optional stopping theorems for sub-, super-, and backward martingales also. For example, if X_n is a nonnegative supermartingale and T is a stopping time for X_n , then $\mathbf{E}X_T \leq \mathbf{E}X_0$. We do not pursue these matters further here.

5.8 The Law of Averages

Suppose that, as a result of some experiment, the event A occurs with probability p (or the event A^c occurs with probability $1-p$). Typically, A might be an event such as “the patient was cured,” “the dart hit the target,” or “the molecule split.” Let N_n be the number of times A occurs in n independent repetitions of this experiment. Now we have shown that N_n is a binomial random variable, and in Example 4.17, we proved that for $\epsilon > 0$

$$(1) \quad \mathbf{P}\left(\left|\frac{1}{n}N_n - p\right| > \epsilon\right) \leq 2 \exp(-n\epsilon^2/4) \\ \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Roughly speaking, this says that the proportion of experiments in which A occurs approaches the probability of A as n increases. (It is pleasing that this agrees with our intuitive notions about events and their probabilities.)

We now develop this simple idea a little. Statements such as (1) are common in probability, so we make a formal definition.

(2) **Definition** Let $(X_n; n \geq 1)$ be a sequence of random variables. We say the sequence X_n converges in probability to X if, for any $\epsilon > 0$, as $n \rightarrow \infty$

$$(3) \quad \mathbf{P}(|X_n - X| > \epsilon) \rightarrow 0.$$

For brevity, this is often written as $X_n \xrightarrow{\text{P}} X$. ▲

In this notation, (1) becomes

$$(4) \quad \frac{1}{n}N_n \xrightarrow{\text{P}} p.$$

Here the limit p is a constant random variable of course.

In Section 5.5 above, we observed that we could write

$$N_n = S_n = \sum_{k=1}^n I_k,$$

where I_k is the indicator of the event that A occurs in the k th experiment, and $\mathbf{E}(I_k) = p$. In the notation of Definition 2, we can thus write (4) as

$$(5) \quad \frac{1}{n}S_n \xrightarrow{\text{P}} \mathbf{E}(I_1).$$

It is natural to wonder whether this result may also hold for sequences other than indicators. The following celebrated result shows that in many cases it does.

(6) **Theorem: Weak Law of Large Numbers** Let $(X_n; n \geq 1)$ be a sequence of independent random variables having the same finite mean and variance, $\mu = \mathbf{E}(X_1)$ and $\sigma^2 = \text{var}(X_1)$. Then, as $n \rightarrow \infty$,

$$(7) \quad \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{\text{P}} \mu.$$

It is customary to write $S_n = \sum_{i=1}^n X_i$ for the partial sums of the X_i .

Proof Recall Chebyshev's inequality: for any random variable Y and $\epsilon > 0$, $\mathbf{P}(|Y| > \epsilon) \leq \mathbf{E}(Y^2)\epsilon^{-2}$. Hence, letting $Y = n^{-1}(S_n - n\mu)$, we have

$$\begin{aligned} \mathbf{P}\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) &\leq \frac{1}{n^2\epsilon^2}\mathbf{E}\left(\left[\sum_{i=1}^n(X_i - \mu)\right]^2\right) \\ &= n^{-2}\epsilon^{-2}\sum_{i=1}^n \text{var}(X_i) = \sigma^2/(n\epsilon^2) \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty, \end{aligned}$$

and (6) is proved. ■

Actually, when S_n is a binomial random variable, we have already shown in Exercise 4.18.3 that more can be said.

- (8) **Theorem** If $(X_i; i \geq 1)$ are independent indicator random variables with $\mathbf{E}(X_i) = p$, then as $n \rightarrow \infty$, for $0 < \epsilon < 1$,

$$\mathbf{P}\left(\left|\frac{1}{m}S_m - p\right| > \epsilon \quad \text{for any } m \geq n\right) \rightarrow 0.$$

Proof Remember that for any events $(A_i; i \geq 1)$, we have

$$\mathbf{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbf{P}(A_i).$$

It follows [using (1)] that

$$\begin{aligned} \mathbf{P}\left(\left|\frac{1}{m}S_m - p\right| > \epsilon \quad \text{for any } m \geq n\right) &\leq \sum_{m=n}^{\infty} \mathbf{P}\left(\left|\frac{1}{m}S_m - p\right| > \epsilon\right) \leq \sum_{m=n}^{\infty} 2e^{-m\epsilon^2/4} \\ &= 2e^{-n\epsilon^2/4}(1 - e^{-\epsilon^2/4})^{-1} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Roughly speaking, this says that not only does the chance of finding $n^{-1}S_n$ far from p vanish, but also the chance that any of $(m^{-1}S_m; m \geq n)$ are far from p vanishes as $n \rightarrow \infty$. Results of this type are called strong laws of large numbers.

Results like those above go some way toward justifying our intuitive feelings about averages in the long run. Such laws of large numbers also have other applications; the following two examples are typical.

- (9) **Example: Monte Carlo Integration** Suppose $f(x)$ is a nonnegative function, and we require $I = \int_a^b f(x) dx$. If f is sufficiently nasty to defy basic methods, a surprisingly effective method of finding I is as follows. Let R be the rectangle $\{x, y : a \leq x \leq b; c \leq y \leq d\}$ where $0 \leq c < f(x) < d$ for $a \leq x < b$. The curve $y = f(x)$ divides R into two disjoint regions, A lying above f and B lying below f .

Now pick a point P_1 at random in R , by which we mean uniformly in R . Then

$$(10) \quad \mathbf{P}(P_1 \in B) = \frac{1}{|R|} \int_a^b f(x) dx = p \quad \text{say.}$$

Now we pick a series $(P_j; j \geq 1)$ of such points independently in R , and let I_j be the indicator of the event that P_j lies in B . Then by the weak law above

$$\frac{1}{n} \sum_{j=1}^n I_j \xrightarrow{P} p,$$

as $n \rightarrow \infty$. Hence, for large n , we may expect that $(|R|/n) \sum_{j=1}^n I_j$ is a reasonable approximation to I and that it improves as n increases.

(We have glossed over one or two details in this simple account; more discussion is provided in Chapter 7.) ●

Here is a similar example.

- (11) **Example: Estimation of Mass Functions** Let $(X_i; i \geq 1)$ be independent and identically distributed with an unknown mass function $f(x)$. Suppose we want to know $f(x)$ for some given x . Let I_k be the indicator of the event that $X_k = x$. Obviously, $\mathbf{E}(I_k) = f(x)$, so by the weak law

$$\frac{1}{n}S_n = \frac{1}{n} \sum_{k=1}^n I_k \xrightarrow{\mathbf{P}} f(x).$$

Thus, $n^{-1}S_n$ should be a good guess at $f(x)$ for large n . ●

Notice that in both these cases, we applied the weak law of large numbers (WLLN) to a binomial random variable arising as a sum of indicators. Despite its simplicity, this is an important special case, so we return to Theorem 6 and note that the proof shows something a little stronger than the WLLN. In fact, in using Chebyshev's inequality we showed that, as $n \rightarrow \infty$,

$$(12) \quad \mathbf{E}\left(\left|\frac{S_n}{n} - \mu\right|^2\right) \rightarrow 0.$$

This type of statement is also widespread in probability and warrants a formal emphasis.

- (13) **Definition** Let X_1, X_2, \dots be a sequence of random variables. If there is a random variable X such that

$$\lim_{n \rightarrow \infty} \mathbf{E}(X_n - X)^2 = 0,$$

then X_n is said to converge in mean square to X . We sometimes write this as $X_n \xrightarrow{\text{m.s.}} X$. ▲

5.9 Convergence

Although important, this section may be omitted at a first reading.

In the preceding section and at various earlier times, we introduced several ideas about the long run behaviour of random variables and their distributions. This seems an appropriate moment to point out the connections between these concepts.

First, we recall our earlier definitions. Here $(X_n; n \geq 1)$ is a sequence of random variables with corresponding distributions $(F_n(x); n \geq 1)$. Also, $F(x)$ is the distribution of a random variable X . Then, as $n \rightarrow \infty$, we say:

- (1) X_n converges in distribution if $F_n(x) \rightarrow F(x)$ whenever $F(x)$ is continuous.
- (2) X_n converges in probability if $\mathbf{P}(|X_n - X| > \epsilon) \rightarrow 0$, for any $\epsilon > 0$.
- (3) X_n converges in mean square if $\mathbf{E}(|X_n - X|^2) \rightarrow 0$, where $\mathbf{E}(X_n^2)\mathbf{E}(X^2) < \infty$.

These are clearly not equivalent statements. For example, let X be integer valued and symmetrically distributed about zero, then let $Y = -X$. By symmetry X and Y have the same distribution, so $|F_Y(x) - F_X(x)| = 0$. However, $\mathbf{P}(|X - Y| > \epsilon) = \mathbf{P}(2|X| > \epsilon) = 1 - \mathbf{P}(X = 0)$. Hence, convergence in distribution does not necessarily imply convergence in probability. This in turn does not necessarily imply convergence in mean square because (2) may hold for random variables without a variance.

What we can say is the following.

Theorem Let $(X_n; n \geq 1)$ be a sequence of random variables having corresponding distributions $(F_n(x); n \geq 1)$. We have the following two results:

- (i) If $X_n \xrightarrow{\text{m.s.}} X$, then $X_n \xrightarrow{\text{P}} X$.
- (ii) If $X_n \xrightarrow{\text{P}} X$, then $F_n(x) \rightarrow F(x)$ at all points x where $F(x)$ is continuous.

Proof (i) By Chebyshev's inequality, for $\epsilon > 0$

$$\mathbf{P}(|X_n - X| > \epsilon) \leq \epsilon^{-2} \mathbf{E}(X_n - X)^2 \rightarrow 0 \quad \text{by hypothesis.}$$

(ii) For any $\epsilon > 0$,

$$\begin{aligned} F_n(x) &= \mathbf{P}(X_n \leq x) \\ &= \mathbf{P}(X_n \leq x, X \leq x + \epsilon) + \mathbf{P}(X_n \leq x, X > x + \epsilon) \\ &\leq F(x + \epsilon) + \mathbf{P}(|X_n - X| > \epsilon). \end{aligned}$$

Likewise,

$$\begin{aligned} F(x - \epsilon) &= \mathbf{P}(X \leq x - \epsilon, X_n \leq x) + \mathbf{P}(X \leq x - \epsilon, X_n > x) \\ &\leq F_n(x) + \mathbf{P}(|X_n - X| > \epsilon). \end{aligned}$$

Hence,

$$F(x - \epsilon) - \mathbf{P}(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x + \epsilon) + \mathbf{P}(|X_n - X| > \epsilon).$$

Now allowing $n \rightarrow \infty$ and $\epsilon \rightarrow 0$ yields the result. ■

Furthermore, it is not always necessary to postulate the existence of a limit random variable X . A typical result is this, which we give without proof.

- (4) **Theorem** Let $(X_n; n \geq 0)$ be a sequence of random variables with finite variance such that, as $j \rightarrow \infty$ and $k \rightarrow \infty$,

$$\mathbf{E}(X_k - X_j)^2 \rightarrow 0.$$

Then there exists a random variable X such that $\mathbf{E}X^2 < \infty$ and $X_n \xrightarrow{\text{m.s.}} X$. Such sequences are called "Cauchy-convergent."

In fact, we can go a little further here, by using Markov's inequality (4.6.2) in this form:

- (5)
$$\mathbf{P}(|X_n - X| > \epsilon) \leq \mathbf{E}|X_n - X|/\epsilon, \quad \text{where } \epsilon > 0.$$

It follows that, as $n \rightarrow \infty$,

$$(6) \quad X_n \xrightarrow{P} X, \quad \text{if } \mathbf{E}|X_n - X| \rightarrow 0.$$

The converse implications to the above result are false in general, but can become true if appropriate extra conditions are added. We single out one famous and important example.

(7) **Theorem: Dominated Convergence** *Let $(X_n; n \geq 0)$ be a sequence of random variables such that $|X_n| \leq Z$ for all n , where $\mathbf{E}Z < \infty$. If $X_n \xrightarrow{P} X$, then $\mathbf{E}|X_n - X| \rightarrow 0$, as $n \rightarrow \infty$.*

Proof Note that because we are dealing with random variables, many statements should include the modifier “with probability 1.” This quickly becomes boring, so we omit this refinement.

Let $Z_n = |X_n - X|$. Because $|X_n| \leq Z$ for all n , it follows that $|X| \leq Z$. Hence,

$$|Z_n| \leq 2Z.$$

Introduce the indicator function $I(A)$, which takes the value 1 if the event A occurs, and is otherwise 0. Then, for $\varepsilon > 0$,

$$(8) \quad \begin{aligned} \mathbf{E}[Z_n] &= \mathbf{E}[Z_n I(Z_n \leq \varepsilon)] + \mathbf{E}[Z_n I(Z_n > \varepsilon)] \\ &\leq \varepsilon + 2\mathbf{E}[ZI(Z_n > \varepsilon)]. \end{aligned}$$

Because $\mathbf{E}|Z| < \infty$ and $Z_n \xrightarrow{P} 0$, the last term decreases to zero as $n \rightarrow \infty$. To see this set,

$$(9) \quad \begin{aligned} \mathbf{E}[ZI(Z_n > \varepsilon)] &= \mathbf{E}[ZI(Z_n > \varepsilon, Z > y)] + \mathbf{E}[ZI(Z_n > \varepsilon, Z \leq y)] \\ &\leq \mathbf{E}[ZI(Z > y)] + y\mathbf{P}(Z_n > \varepsilon). \end{aligned}$$

Now choose first y and then n as large as we please, to make the right side arbitrarily small. The result follows because ε was arbitrary. ■

If Z is a finite constant, then this yields a special case called the *bounded convergence theorem*. There is another theorem called the *monotone convergence theorem*, which asserts that

$$(10) \quad \mathbf{E}X_n \rightarrow \mathbf{E}X \quad \text{if } X_n \uparrow X, \quad \text{as } n \rightarrow \infty.$$

We offer no proof of this, though we will use it as necessary.

Now we can use Theorem (7) immediately to prove some optional stopping theorems for martingales.

(11) **Theorem** *Let X_n be a martingale and T a stopping time for X_n . Then $\mathbf{E}X_T = \mathbf{E}X_0$ if any of the following hold for some real positive finite constant K :*

- (a) T is bounded (i.e., $T \leq K < \infty$).
- (b) $|X_n| \leq K$ for all n , and $\mathbf{P}(T < \infty) = 1$.
- (c) $\mathbf{E}T < \infty$ and $\mathbf{E}(|X_{n+1} - X_n| | X_0, \dots, X_n) \leq K, n < T$.
- (d) $\mathbf{E}|X_T| < \infty, \mathbf{P}(T < \infty) = 1$, and $\mathbf{E}(X_n I(T > n)) \rightarrow 0$ as $n \rightarrow \infty$.

Remark In practice a popular technique when $\mathbf{P}(T < \infty) = 1$ is to apply part (a) of the theorem at $T \wedge n$, and then let $n \rightarrow \infty$, using Dominated or Monotone convergence to obtain the required result.

Proof We proved this when (a) holds in (5.7.14), where we also showed that

$$\mathbf{E}(X_{T \wedge n} - X_0) = 0.$$

Allowing $n \rightarrow \infty$, we have $X_{T \wedge n} - X_0 \rightarrow X_T - X_0$. Now suppose (b) holds. Because the sequence is bounded, using (7) above gives the result. Next, we suppose (c) holds and write

$$|X_{T \wedge n} - X_0| = \left| \sum_{r=1}^{T \wedge n} (X_r - X_{r-1}) \right| \leq \sum_{r=1}^{\infty} |X_r - X_{r-1}| I(T \geq r).$$

Hence, because (c) holds,

$$\mathbf{E} \sum_{r=1}^{\infty} |X_r - X_{r-1}| I(T \geq r) \leq \sum_{r=1}^{\infty} K \mathbf{P}(T \geq r) = K \mathbf{E}T.$$

Because $\mathbf{E}T < \infty$, we can use the dominated convergence theorem as $n \rightarrow \infty$ to obtain the required result. For (d) see Problem 5.44 ■

Similar results hold for submartingales and supermartingales, and are proved in the same way; we omit the details.

We conclude with another important result. One of the principal features of martingales is that they converge with only weak additional constraints on their nature. Here is one example.

(12) **Example: Martingale Convergence** Let $(Y_n; n \geq 0)$ be a martingale such that $\mathbf{E}Y_n^2 \leq K$ for all n . Show that Y_n converges in mean square as $n \rightarrow \infty$.

Solution For $r \geq i$, $\mathbf{E}(Y_r | Y_0, \dots, Y_i) = \mathbf{E}(\mathbf{E}(Y_r | Y_{r-1}, \dots, Y_0) | Y_0, \dots, Y_i) = \mathbf{E}(Y_{r-1} | Y_0, \dots, Y_i)$. Iterating shows that $\mathbf{E}(Y_r | Y_0, \dots, Y_i) = Y_i$. Hence,

$$\mathbf{E}(Y_r Y_i) = \mathbf{E}[\mathbf{E}(Y_r Y_i | Y_0, \dots, Y_i)] = \mathbf{E}(Y_i^2).$$

It follows that for $i \leq j \leq k$

$$(13) \quad \mathbf{E}((Y_k - Y_j)Y_i) = \mathbf{E}(Y_k Y_i) - \mathbf{E}(Y_j Y_i) = 0.$$

Thus, after some algebra,

$$\mathbf{E}[(Y_k - Y_j)^2 | Y_0, \dots, Y_i] = \mathbf{E}(Y_k^2 | Y_0, \dots, Y_i) - \mathbf{E}(Y_j^2 | Y_0, \dots, Y_i).$$

Therefore, $0 \leq \mathbf{E}(Y_k - Y_j)^2 \leq \mathbf{E}Y_k^2 - \mathbf{E}Y_j^2$. Now $(\mathbf{E}Y_n^2; n \geq 1)$ is nondecreasing and bounded, and therefore converges. Finally, we deduce that $\mathbf{E}(Y_k - Y_j)^2 \rightarrow 0$ as $k, j \rightarrow \infty$, and this shows $Y_k \xrightarrow{\text{ms}} Y$; by (4). ●

The above example included a proof of the following small but important result.

- (14) **Corollary: Orthogonal Increments** If $(X_n; n \geq 0)$ is a martingale and $i \leq j \leq k \leq m$, then using (11)

$$\mathbf{E}[(X_m - X_k)(X_j - X_i)] = 0,$$

which is called the orthogonal increments property, because of (5.3.9) ■

5.10 Review and Checklist for Chapter 5

This chapter extended ideas from earlier chapters to enable us to make probabilistic statements about collections and sequences of random variables. The principal instrument to help us is the joint probability distribution and joint probability mass function of discrete random variables. We defined the concepts of independence and conditioning for random variables. Jointly distributed random variables have joint moments, and we looked at covariance and correlation. Conditional expectation is a concept of great importance and utility; we used all these ideas in examining the properties of functions of random variables. Finally, we looked at random walks, martingales, stopping times, optional stopping, sequences of random variables, and simple ideas about their convergence.

We give details of these principal properties for pairs of random variables; all these expressions are easily generalized to arbitrary collections, at the expense of more notation and space.

SYNOPSIS OF FORMULAE:

Pairs X and Y of such variables have a joint mass function

$$f(x, y) = \mathbf{P}(X = x, Y = y),$$

which appears in the

Key Rule for joint distributions: for any set C of possible values of (X, Y)

$$\mathbf{P}((X, Y) \in C) = \sum_{(x,y) \in C} f(x, y).$$

In particular, we have the

Joint distribution:

$$F(x, y) = \mathbf{P}(X \leq x, Y \leq y) = \sum_{\substack{u \leq x \\ v \leq y}} f(u, v).$$

Marginals:

$$f_X(x) = \sum_y f(x, y), \quad f_Y(y) = \sum_x f(x, y).$$

Functions:

$$\mathbf{P}(g(X, Y) = z) = \sum_{x,y: g=z} f(x, y).$$

Sums:

$$\mathbf{P}(X + Y = z) = \sum_x f(x, z - x) = \sum_y f(z - y, y).$$

Independence: X and Y are independent if $f(x, y) = f_X(x)f_Y(y)$ for all x and y .

Conditioning: The conditional probability mass function of X given Y is

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}, \quad f_Y(y) > 0.$$

The Key Rule for conditional mass functions:

$$\mathbf{P}(X \in A | Y = y) = \sum_{x \in A} f_{X|Y}(x|y),$$

and the **Partition Rule** says

$$f_X(x) = \sum_y f_Y(y)f_{X|Y}(x|y).$$

Expectation: The expected value of the random variable $g(X, Y)$ is

$$\mathbf{E}g(X, Y) = \sum_{x,y} g(x, y)f(x, y).$$

In particular, for constants a and b ,

$$\mathbf{E}(ag(X, Y) + bh(X, Y)) = a\mathbf{E}g + b\mathbf{E}h.$$

Moments: The covariance of X and Y is

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]$$

and the correlation coefficient is

$$\rho(X, Y) = \text{cov}(X, Y) / \{\text{var } X \text{ var } Y\}^{1/2}.$$

We note that

$$\text{cov}\left(\sum X_i, \sum Y_j\right) = \sum_{i,j} \text{cov}(X_i, Y_j).$$

Independence: When X and Y are independent $\mathbf{E}(XY) = \mathbf{E}X\mathbf{E}Y$, so that X and Y are uncorrelated and

$$\text{cov}(X, Y) = \rho(X, Y) = 0.$$

In this case, when the X_i are independent,

$$\text{var}\left(\sum X_i\right) = \sum \text{var } X_i.$$

Conditional expectation: The conditional expectation of X given $Y = y$ is

$$\mathbf{E}(X|Y = y) = \sum_x xf_{X|Y}(x|y) = \sum_x xf(x, y)/f_Y(y).$$

For any pair of random variables where both sides exist,

$$\mathbf{E}[\mathbf{E}(X|Y)] = \mathbf{E}X.$$

Key properties: the conditional expectation $\mathbf{E}(X|Y)$ satisfies

$$\mathbf{E}(X|Y) = \mathbf{E}X, \quad \text{if } X \text{ and } Y \text{ are independent.}$$

$$\mathbf{E}(Xg(Y)|Y) = g(Y)\mathbf{E}(X|Y), \quad \text{the pull-through property.}$$

$$\mathbf{E}(\mathbf{E}(X|Y); Z|Y) = \mathbf{E}(X|Y), \quad \text{the tower property.}$$

Conditional variance:

$$\text{var } X = \text{var } \mathbf{E}(X|Y) + \mathbf{E}\text{var}(X|Y),$$

where

$$\text{var}(X|Y) = \mathbf{E}(X^2|Y) - [\mathbf{E}(X|Y)]^2.$$

Remark Experience of student calculations leads us to stress that it is *not* in general true that

$$\text{var } X = \mathbf{E} \text{var}(X|Y).$$

Conditional independence: X and Y are conditionally independent given $Z = z$ if, for all x and y ,

$$\mathbf{P}(X = x, Y = y|Z = z) = f_{X|Z}(x|z)f_{Y|Z}(y|z).$$

Checklist of Terms for Chapter 5

- 5.1 joint probability mass function
 - marginal mass function
- 5.2 independent random variables
- 5.3 expectation
 - covariance
 - joint moments
 - orthogonal random variables
 - correlation coefficient
 - Cauchy–Schwarz inequality
- 5.4 sums of indicators
 - inclusion–exclusion inequalities
 - convolution
- 5.5 conditional mass function
 - conditional expectation
 - random sum
 - tower property
 - discrete partition rule
- 5.6 simple random walk
 - first passage time
 - recurrence time

hitting time theorem
 reflection principle
 ballot theorem

5.7 martingale

stopping time
 optional stopping
 Wald's equation

5.8 weak law of large numbers

convergence mean square

5.9 convergence of distributions

dominated convergence
 martingale convergence

WORKED EXAMPLES AND EXERCISES

5.11 Example: Golf

Arnold and Bobby play a complete round of 18 holes at golf. Holes are independent, and any hole is won by Arnold with probability p , won by Bobby with probability q , or it is halved with probability r . Of the 18 Arnold wins X , Bobby wins Y , and Z are halved.

- (a) Find the joint mass function of X , Y , and Z .
 (b) What is the marginal distribution of X ?
 (c) Show that the correlation coefficient of X and Y is

$$\rho(X, Y) = - \left(\frac{pq}{(1-p)(1-q)} \right)^{\frac{1}{2}}.$$

Solution Let A_k be the event that Arnold wins the k th hole, B_k the event that he loses the k th hole, and H_k the event that the k th hole is halved. We illustrate the possibilities by giving more than one solution.

(a) **I** A typical outcome is a string of x A s, y B s, and z H s. Such a sequence has probability $p^x q^y r^z$ of occurring, by independence, and by Theorem 3.3.2 there are $18!/(x!y!z!)$ such sequences. Hence,

$$(1) \quad \mathbf{P}(X = x, Y = y, Z = z) = \frac{18! p^x q^y r^z}{x! y! z!}; \quad x + y + z = 18.$$

II By Definition 5.5.1,

$$\mathbf{P}(X = x, Y = y, Z = z) = \mathbf{P}(X = x, Y = y | Z = z) \mathbf{P}(Z = z).$$

Now the number of holes halved is just the number of successes in 18 Bernoulli trials with $\mathbf{P}(\text{success}) = r$. Hence, by Example 5.4.1 (or Example 4.2.3),

$$\mathbf{P}(Z = z) = r^z (1 - r)^{18-z} \binom{18}{z}.$$

Now for any given hole, $\mathbf{P}(A|H^c) = p/(p + q)$. Hence, given $Z = z$, the number of holes won by Arnold is just the number of successes in $18 - z$ Bernoulli trials with $\mathbf{P}(\text{success}) = p/(p + q)$. Therefore,

$$\mathbf{P}(X = x, Y = y|Z = z) = \binom{18 - z}{x} \left(\frac{p}{p + q}\right)^x \left(\frac{q}{p + q}\right)^y,$$

where $x + y = 18 - z$. Thus,

$$\begin{aligned} \mathbf{P}(X = x, Y = y, Z = z) &= \frac{(x + y)!}{x!y!} \frac{p^x q^y}{(1 - r)^{x+y}} r^z (1 - r)^{18-z} \frac{18!}{(x + y)!z!} \\ &= \frac{18! p^x q^y r^z}{x!y!z!}. \end{aligned}$$

(b) **I** As in (5.1.5), we have

$$\begin{aligned} (2) \quad \mathbf{P}(X = x) &= \sum_{y,z} \mathbf{P}(X = x, Y = y, Z = z) \\ &= \sum_{y=0}^{18-x} \frac{18! p^x}{x!(18-x)!} \frac{(18-x)!}{y!(18-x-y)!} q^y r^{18-x-y} \\ &= \binom{18}{x} p^x (q + r)^{18-x}, \end{aligned}$$

which is binomial with parameters 18 and p .

II Either Arnold succeeds with probability p in winning each hole, or he fails with probability $1 - p = q + r$. Hence, by Example 5.4.1, the mass function of X is binomial as in (2).

(c) Let I_k be the indicator of the event A_k that Arnold wins the k th hole, and J_k the indicator of the event B_k that Bobby wins it. Then $I_k J_k = 0$, and I_j is independent of J_k for $j \neq k$. Hence, $\mathbf{E}(I_j I_k) = pq$ for $j \neq k$, and therefore,

$$\mathbf{E}(XY) = \mathbf{E}\left(\sum_{k=1}^{18} I_k \sum_{j=1}^{18} J_j\right) = 18 \times 17pq.$$

Thus,

$$\text{cov}(X, Y) = 18 \times 17pq - 18p \times 18q = -18pq.$$

Finally, we note that because X and Y are binomial, we have $\text{var}(X) = 18p(1 - p)$ and $\text{var}(Y) = 18q(1 - q)$. Therefore,

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{(\text{var}(X) \text{var}(Y))^{1/2}} = \frac{-18pq}{(18p(1 - p) \cdot 18q(1 - q))^{1/2}},$$

as required.

- (3) **Exercise** What is $\rho(Y, Z)$?
- (4) **Exercise** What is the conditional mass function of X , given $X + Y = m$?
- (5) **Exercise** What is the probability that the match is halved?
- (6) **Exercise** What is $\mathbf{E}(X|Y)$?
- (7) **Exercise** What is $\mathbf{E}(X|Y, Z)$?

5.12 Example: Joint Lives

Suppose that $2m$ individuals constitute m married couples at some given initial date. We want to consider the survivors at some given later date. Suppose that each individual is alive at the later date with probability p independently of the others. Let A be the number of individuals then alive, and let S be the number of surviving couples in which both the partners are alive. Show that

$$\mathbf{E}(S|A) = \frac{A(A-1)}{2(2m-1)}.$$

Remark This problem was discussed by Daniel Bernoulli in 1768.

Solution Let S_a be the number of surviving couples given that $A = a$. We give several methods of solution; you can choose your favourite, or of course find a better one.

I Let I_j be the indicator of the event that the j th couple survives. Then

$$\mathbf{E}(S_a) = \mathbf{E}\left(\sum_1^m I_j\right) = m\mathbf{E}(I_1) = m\mathbf{P}(I_1 = 1)$$

because the chance of survival is the same for every couple. Now we can choose the a survivors in $\binom{2m}{a}$ ways, and the number of these in which the first couple remain alive is $\binom{2m-2}{a-2}$. (This is the number ways of choosing $a-2$ other survivors from the other $m-1$ couples.)

Because these are equally likely outcomes

$$\mathbf{P}(I_1 = 1) = \binom{2m-2}{a-2} / \binom{2m}{a} = \frac{a(a-1)}{2m(2m-1)}.$$

Hence, $\mathbf{E}(S_a) = a(a-1)/(2(2m-1))$.

II Suppose that the a individuals remaining alive include x couples. If one more individual were to die, then the expected number of couples remaining would be $\mathbf{E}(S_{a-1}|S_a = x)$. To evaluate this, observe that if a widow/er dies then there are still x couples; however, if a survivor's spouse dies, there are now $x-1$ couples. The probability of a widow/er's death is $(a-2x)/a$; the probability of the death of one individual of the x couples is $(2x/a)$. Hence,

$$\mathbf{E}(S_{a-1}|S_a = x) = \frac{x(a-2x)}{a} + \frac{(x-1)2x}{a} = \frac{(a-2)}{a}x.$$

Hence, by Theorem 5.5.6,

$$\mathbf{E}(S_{a-1}) = \mathbf{E}(\mathbf{E}(S_{a-1}|S_a)) = \frac{a-2}{a}\mathbf{E}(S_a).$$

This relation may be iterated on the left or the right, to give either

$$\mathbf{E}(S_a) = \frac{a(a-1)}{2}\mathbf{E}(S_2) = \frac{a(a-1)}{2(2m-1)}$$

or

$$\mathbf{E}(S_a) = \frac{a(a-1)}{2m(2m-1)}\mathbf{E}(S_{2m}) = \frac{a(a-1)}{2(2m-1)}.$$

III Number the couples $1, \dots, m$. Let Y_i be the indicator of the event that the male of the i th couple survives and X_i the indicator of the event that the female of the i th couple survives. Then $S_a = \sum_1^m X_i Y_i$.

Now $\mathbf{P}(Y_1 = 1|X_1 = 1) = (a-1)/(2m-1)$ and $\mathbf{P}(X_1 = 1) = a/(2m)$. Hence,

$$\mathbf{E}(S_a) = \sum_1^m \mathbf{E}(Y_i X_i) = m\mathbf{E}(Y_1|X_1 = 1)\mathbf{P}(X_1 = 1) = m \frac{a}{2m} \frac{a-1}{2m-1},$$

as required.

- (1) **Exercise** Find the mass function of S , and write down its mean and variance.
- (2) **Exercise** Show that $\mathbf{E}(AS) = 2m((m-1)p^3 + p^2)$.
- (3) **Exercise** Show that $\mathbf{E}(A|S) = 2mp + 2(1-p)S$.
- (4) **Exercise** Show that the correlation $\rho(A, S)$ is given by $\rho(A, S) = (2p/(1+p))^{\frac{1}{2}}$.
- (5) **Exercise** Suppose that males and females have different death rates, so the probability of a male surviving is μ and the probability of a female surviving is ϕ . Show that S has a $B(m, \mu\phi)$ mass function. What is the mass function of A ? What is $\mathbf{E}(A)$?
- (6) **Exercise** When males and females have different survival rates μ and ϕ , find $\mathbf{E}(A|S)$ and hence show that in this case

$$\rho(A, S) = \frac{(2 - \phi - \mu)(\phi\mu)^{\frac{1}{2}}}{((1 - \phi\mu)(\phi(1 - \phi) + \mu(1 - \mu)))^{\frac{1}{2}}}.$$

5.13 Example: Tournament

Suppose that 2^n tennis players enter a knock-out singles tournament, and the players are completely ranked (with no ties). The draw for the tournament is at random, and we suppose that in any match the higher ranked player always wins. Let R_n be the rank of the losing finalist; find $\mathbf{E}(R_n)$, and show that as $n \rightarrow \infty$

$$\mathbf{E}(R_n) \rightarrow 3.$$

Solution The losing finalist comes from the half of the draw not containing the top-ranked player. These 2^{n-1} players have ranks $N_1 < N_2 < \dots < N_{2^{n-1}}$, which are drawn at random from the $2^n - 1$ integers $\{2, 3, 4, \dots, 2^n\}$, and $R_n = N_1$.

Let X_1, X_2, \dots be the numbers of players drawn with the top-ranked player, between successive players drawn for the other half. That is to say

$$\begin{aligned} X_1 &= N_1 - 2 \\ X_k &= N_k - N_{k-1} - 1, \quad 2 \leq k \leq 2^{n-1}, \\ X_{2^{n-1}+1} &= 2^n - N_{2^{n-1}}. \end{aligned}$$

By symmetry, for all j and k , $\mathbf{E}(X_k) = \mathbf{E}(X_j)$. Hence,

$$(1) \quad (2^{n-1} + 1)\mathbf{E}(X_1) = \mathbf{E}\left(\sum_1^{2^{n-1}+1} X_k\right) = 2^n - 1 - 2^{n-1}.$$

Thus,

$$\begin{aligned} \mathbf{E}(R_n) &= \mathbf{E}(N_1) = \mathbf{E}(X_1) + 2 \\ &= \frac{2^{n+1} - 2^{n-1} + 1}{2^{n-1} + 1} \rightarrow 3, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

- (2) **Exercise** Suppose that the ranking allows ties (so that, for example, a possible ranking is 1, 1, 3, 3, 3, 6, ...). Show that as $n \rightarrow \infty$, $\lim \mathbf{E}(R_n) \leq 3$.
- (3) **Exercise** Suppose that there are $3 \times 2^{n-1}$ entrants, and these are divided at random into a group of size 2^{n-1} and a group of size 2^n who then knock each other out in the usual way to provide two finalists. Find $\mathbf{E}(R_n)$ and show that $\mathbf{E}(R_n) \rightarrow \frac{7}{2}$.
- (4) **Exercise** 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 $(b+r+1)/(b+1)$.
The balls are replaced, and then removed at random until all the balls remaining are of the same colour. Show that the expected number remaining is $r/(b+1) + b/(r+1)$. What is the probability p_r that they are all red?
- (5) **Exercise** Let X_1, X_2, \dots be independent and identically distributed. What is

$$\mathbf{E}\left(\frac{\sum_1^m X_i}{\sum_1^n X_i}\right) \text{ when } m \leq n?$$

5.14 Example: Congregations

Suppose that n initially separate congregations of people are then united, and one person is picked at random from the united group. Let the size of the congregation of which she was originally a member be Y . If the respective sizes of the original congregations are the random variables $(X_i; 1 \leq i \leq n)$, show that

$$(1) \quad \mathbf{E}(Y) = \mathbf{E}\left(\frac{\sum_1^n X_i^2}{\sum_1^n X_i}\right)$$

$$(2) \quad \geq \frac{1}{n} \sum_1^n \mathbf{E}(X_i).$$

Solution Let us use conditional expectation. Given that $X_i = x_i$ for $1 \leq i \leq n$, the probability that the selected individual was in the r th congregation initially is $x_r / (\sum_1^n x_i)$.

Hence,

$$\mathbf{E}(Y|X_1 = x_1, \dots, X_n = x_n) = \sum_1^n x_r \mathbf{P}(Y = x_r) = \sum_{r=1}^n x_r^2 / \left(\sum_{i=1}^n x_i \right).$$

Therefore,

$$\mathbf{E}(Y) = \mathbf{E}(\mathbf{E}(Y|X_1, \dots, X_n)) = \mathbf{E} \left(\frac{\sum_1^n X_i^2}{\sum_1^n X_i} \right).$$

Now recall Cauchy’s inequality for real numbers $(x_i; 1 \leq i \leq n)$ and $(y_i; 1 \leq i \leq n)$, namely,

$$(3) \quad \left(\sum_1^n x_i y_i \right)^2 \leq \left(\sum_1^n x_i^2 \right) \left(\sum_1^n y_i^2 \right).$$

Setting $y_i = 1$, for all i , yields $(\sum_1^n x_i^2)/(\sum_1^n x_i) \geq \sum_1^n x_i/n$, and the required inequality (2) follows.

Remark Observe that if a congregation is picked at random by choosing a number in $\{1, 2, \dots, n\}$ at random, then the expected size of the chosen congregation is $\frac{1}{n} \sum_1^n \mathbf{E}(X_i)$. The fact that a member picked at random was in a larger expected congregation is a form of sampling “paradox.”

- (4) **Exercise** For what distributions of X_i , if any, does the expected size of a randomly selected individual’s group actually equal the mean size of groups?
- (5) **Exercise** Family sizes are independent and identically distributed with mean μ . If you pick an individual at random find the probability that she is the k th born of her family, and the expectation of her order of birth in her family. Compare this with μ .
- (6) **Exercise** Use the Cauchy–Schwarz inequality Lemma 5.3.13 to prove Cauchy’s inequality (3).

5.15 Example: Propagation

A plant sheds N seeds, where N is a binomial random variable with parameters n and p . Each seed germinates with probability γ independently of all the others. Let S denote the number of resulting seedlings. Find:

- (a) The conditional mass function of S given N .
- (b) The joint mass function of S and N .
- (c) The probability mass function of S .
- (d) The conditional mass function of N given S .

Solution (a) Given that there are i seeds, that is $N = i$, the germination of any one can be regarded as a Bernoulli trial with $\mathbf{P}(\text{success}) = \gamma$. Then by Example 5.4.1 (also discussed in Example 4.2.3), the total number of successes is a binomial random variable

with parameters i and γ . So

$$\mathbf{P}(S = j|N = i) = \binom{i}{j} \gamma^j (1 - \gamma)^{i-j}; \quad 0 \leq j \leq i.$$

(b) Now for the joint mass function

$$\begin{aligned} \mathbf{P}(N = i \cap S = j) &= \mathbf{P}(S = j|N = i)\mathbf{P}(N = i) \\ &= \binom{i}{j} \gamma^j (1 - \gamma)^{i-j} \binom{n}{i} p^i (1 - p)^{n-i}. \end{aligned}$$

(c) Now we require the marginal mass function of S , which is given by

$$\begin{aligned} \mathbf{P}(S = j) &= \sum_{i=j}^n \mathbf{P}(N = i, S = j) \quad \text{using (5.1.6)} \\ &= \sum_{i=j}^n \frac{\gamma^j (1 - \gamma)^{i-j} p^i (1 - p)^{n-i}}{(n - i)! j! (i - j)!} \\ &= (1 - p)^n \left(\frac{\gamma}{1 - \gamma} \right)^j \binom{n}{j} \sum_{i=j}^n \left(\frac{(1 - \gamma)p}{1 - p} \right)^i \frac{(n - j)!}{(n - i)! (i - j)!} \\ &= \binom{n}{j} (\gamma p)^j (1 - \gamma p)^{n-j}. \end{aligned}$$

Thus, S is binomial with parameters n and γp .

(d) Finally,

$$\begin{aligned} \mathbf{P}(N = i|S = j) &= \frac{\mathbf{P}(N = i \cap S = j)}{\mathbf{P}(S = j)} = \binom{n - j}{n - i} \left(\frac{1 - p}{1 - \gamma p} \right)^{n-i} \\ &\quad \times \left(1 - \frac{1 - p}{1 - \gamma p} \right)^{i-j}; \quad j \leq i \leq n. \end{aligned}$$

Thus, the variable $N - S$ given that $S = j$ germinate is binomial with parameters $n - j$ and $(p - p\gamma)/(1 - \gamma p)$.

- (1) **Exercise** Find $\mathbf{E}(N|S)$.
- (2) **Exercise** Find $\mathbf{E}(S|N)$.
- (3) **Exercise** Find $\text{cov}(N, S)$ and $\rho(N, S)$.
- (4) **Exercise** Each seedling independently succeeds in growing into a tree with probability τ , or succumbs to wilt with probability $1 - \tau$. Let T be the number of resulting trees.

Find the joint probability mass function of N , S , and T , and also the conditional mass function of N given T .

- (5) **Exercise** Find the joint mass function of N and T given that $S = s$.
- (6) **Exercise** Find the conditional covariance of N and T given that $S = s$.

5.16 Example: Information and Entropy

(a) Let the random variable X take a finite number of values $(x_i; 1 \leq i \leq n)$, with mass function $\mathbf{P}(X = x) = f(x)$. Suppose that $(a_i; 1 \leq i \leq n)$, are such that $a_i > 0$ for $1 \leq i \leq n$.

n and $\sum_{i=1}^n a_i = 1$. Show that

$$-\sum_{i=1}^n f(i) \log a_i \geq -\sum_{i=1}^n f(i) \log f(i)$$

with equality if and only if $a_i = f(i)$ for all i .

(b) The random variables X and Y take a finite number of values and have joint mass function $f(x, y)$. Define

$$I(X, Y) = \sum_x \sum_y f(x, y) \log \left(\frac{f(x, y)}{f_X(x)f_Y(y)} \right).$$

Show that $I \geq 0$, with equality if and only if X and Y are independent.

Solution (a) By definition,

$$\begin{aligned} \log y &= \int_1^y x^{-1} dx \leq \int_1^y dx \quad \text{with equality if } y = 1, \\ &= y - 1. \end{aligned}$$

Hence,

$$(1) \quad \log y \leq y - 1$$

with equality if and only if $y = 1$. Therefore,

$$\begin{aligned} -\sum_i f(i) \log f(i) + \sum_i f(i) \log a_i &= \sum_i f(i) \log \left(\frac{a_i}{f(i)} \right) \\ &\leq \sum_i f(i) \left(\frac{a_i}{f(i)} - 1 \right) \quad \text{by (1)} \\ &= 0, \end{aligned}$$

with equality if and only if $f(i) = a_i$ for all i .

(b) The positive numbers $f_X(x)f_Y(y)$ satisfy $\sum f_X(x)f_Y(y) = 1$. Therefore, by part (a),

$$\sum_{x,y} f(x, y) \log f(x, y) \geq \sum_{x,y} f(x, y) \log(f_X(x)f_Y(y))$$

with equality if and only if for all x and y $f(x, y) = f_X(x)f_Y(y)$. But this is a necessary and sufficient condition for the independence of X and Y .

(2) **Exercise** Show that $I = \mathbf{E}(\log f(X, Y)) - \mathbf{E}(\log f_X(X)) - \mathbf{E}(\log f_Y(Y))$.

(3) **Exercise** Show that if the conditional mass function of X given that $Y = y$ is $f(x|y)$, we have

$$\begin{aligned} I &= \sum_{x,y} f_Y(y)f(x|y) \log f(x|y) - \mathbf{E}(\log f_X(X)) \\ &= \sum_{x,y} f_X(x)f(y|x) \log f(y|x) - \mathbf{E}(\log f_Y(Y)). \end{aligned}$$

(4) **Exercise** A die is rolled twice, yielding the respective scores X and Y . Let $Z = \max\{X, Y\}$. Find $I(X, Z)$ and $I(Z, X)$.

Remark The quantity I is sometimes said to be the information about X conveyed by Y . It is interesting that this is equal to the information about Y conveyed by X .

The quantity $H(X) = \mathbf{E}(-\log f_X(X))$ is known as the entropy or uncertainty of X , and $H(X|Y) = H(X) - I(X, Y)$ is known as the conditional entropy (or uncertainty) of X given Y . It is interpreted as the uncertainty of X , reduced by the information conveyed about X by Y .

- (5) **Exercise** Show that $H(X|X) = 0$.
 (6) **Exercise** Show that if $H(X|Y) = 0 = H(Y|X)$ and $H(Y|Z) = H(Z|Y) = 0$, then $H(X|Z) = H(Z|X) = 0$.

5.17 Example: Cooperation

Achilles and his two friends, Briseis and Chryseis, play a cooperative game. They possess a die with n faces, and each of them rolls it once. Then $I(AC)$ is the indicator of the event that Achilles and Chryseis each turn up the same face of the die. $I(AB)$ and $I(BC)$ are defined similarly. Show that $I(AB)$, $I(AC)$, and $I(BC)$ are pairwise independent if and only if the die is unbiased.

Solution Let the die, when rolled, show its k th face with probability $f(k)$. Then

$$\mathbf{P}(I(AB) = 1, I(BC) = 1) = \mathbf{P}(\text{all three rolls show the same face}) = \sum_{k=1}^n (f(k))^3$$

and

$$\mathbf{P}(I(AB) = 1) = \mathbf{P}(\text{two rolls show the same face}) = \sum_{k=1}^n (f(k))^2.$$

Pairwise independence then requires that

$$(1) \quad \left(\sum_k (f(k))^2 \right)^2 = \sum_k (f(k))^3.$$

Now let X be a random variable that takes the value $f(k)$ with probability $f(k)$. Then (1) states that

$$0 = \mathbf{E}(X^2) - (\mathbf{E}(X))^2 = \mathbf{E}(X - \mathbf{E}(X))^2 = \text{var}(X).$$

Hence, X must be constant by Example 4.6.10, and so the die is unbiased because $f(k) = \frac{1}{n}$; $1 \leq k \leq n$. In this case, it is easy to check that

$$\begin{aligned} \mathbf{P}(I(AB) = 1, I(BC) = 0) &= \sum_{k=1}^n \frac{1}{n} \cdot \frac{n-1}{n} \cdot \frac{1}{n} = \frac{1}{n} \cdot \frac{n-1}{n} \\ &= \mathbf{P}(I(AB) = 1)\mathbf{P}(I(BC) = 0) \end{aligned}$$

and two other conditions are satisfied. The indicators are pairwise independent only in this case.

- (2) **Exercise** Are the indicators independent?
 (3) **Exercise** Find the mean and variance of $Z = I(AB) + I(BC) + I(AC)$.
 (4) **Exercise** If n women play a similar game, and $I(A_i, A_j)$ is the indicator of the event that the i th and j th women turn up the same face of the die, find the mean and variance of $\sum_{i \neq j} I(A_i A_j)$.

5.18 Example: Strange But True

Let $(S_n; n \geq 0)$ be a simple symmetric random walk with $S_0 = 0$. Let $f_0(n) = \mathbf{P}(T_0 = n)$ be the probability that the walk first returns to zero at the n th step, and let $u(n) = \mathbf{P}(S_n = 0)$. Let V_n be the number of values which a walk of n steps has visited exactly once.

(a) Show that

(1)
$$f_0(2k) = u(2k - 2) - u(2k).$$

(b) Deduce that

(2)
$$\mathbf{P}\left(\prod_{k=1}^{2n} S_k \neq 0\right) = \mathbf{P}(T_0 > 2n) = u(2n).$$

(c) Hence, show that for all $n \geq 1$

(3)
$$\mathbf{E}(V_n) = 2.$$

Solution (a) By symmetry and the reflection principle, using Theorem 5.6.17,

$$\begin{aligned} f_0(2k) &= \frac{1}{2k-1} \mathbf{P}(S_{2k-1} = 1) = \frac{2^{-(2k-1)}}{2k-1} \binom{2k-1}{k} \\ &= \frac{2^{-2k}}{2k-1} \binom{2k}{k} = 2^{-2k+2} \binom{2k-2}{k-1} - 2^{-2k} \binom{2k}{k} = u(2k-2) - u(2k). \end{aligned}$$

(b)

$$\begin{aligned} \mathbf{P}(T_0 > 2n) &= 1 - \sum_{k=1}^n f_0(2k) = 1 - \sum_{k=1}^n (u(2k-2) - u(2k)) \quad \text{by (1),} \\ &= u(2n). \end{aligned}$$

(c) Clearly, $V_1 = 2$, so it suffices to show that $\mathbf{E}(V_n) = \mathbf{E}(V_{n-1})$ for $n \geq 2$. Let V'_n be the number of points visited just once by S_1, S_2, \dots, S_n . This has the same distribution as V_{n-1} , and so also the same expectation. Let T_0 be the time of first return to the origin.

Now,

(4)
$$V_n = \begin{cases} V'_n + 1 & \text{if } T_0 > n \\ V'_n - 1 & \text{if } S_1, \dots, S_n \text{ revisits zero exactly once} \\ V'_n & \text{otherwise} \end{cases}$$

Hence,

$$\begin{aligned}
 (5) \quad \mathbf{E}(V_n) - \mathbf{E}(V_{n-1}) &= \mathbf{E}(V_n) - \mathbf{E}(V'_n) \\
 &= \mathbf{P}(T_0 > n) - \mathbf{P}(S_1, \dots, S_n \text{ revisits } 0 \text{ exactly once}) \quad \text{by (4),} \\
 &= \mathbf{P}(T_0 > n) - \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \mathbf{P}(T_0 = 2k) \mathbf{P}\left(\prod_{i=2k+1}^n S_i \neq 0\right) \\
 &= \mathbf{P}(T_0 > n) - \sum_{k=1}^{\lfloor \frac{n}{2} \rfloor} \mathbf{P}(T_0 = 2k) \mathbf{P}(S_{n-2k} = 0) \quad \text{by (2).} \\
 &= \mathbf{P}(T_0 > n) - \mathbf{P}(S_n = 0) \\
 &= 0 \quad \text{by (2).}
 \end{aligned}$$

(6) **Exercise** Show that $2kf_0(2k) = u(2k - 2)$.

(7) **Exercise** Show that $\mathbf{P}(S_{2n} = 0) = (1/2n)\mathbf{E}(|S_{2n}|)$.

(8) **Exercise** Let L_{2n} be the time of the last visit to 0 up to time $2n$. Show that $\mathbf{P}(L_{2n} = 2k) = u(2k)u(2n - 2k)$.

Show that if k and n increase in such a way that $k/n = x$, then

$$\begin{aligned}
 \mathbf{P}\left(\frac{L_{2n}}{2n} \leq x\right) &\rightarrow \frac{2}{\pi} \sin^{-1} x; \quad 0 \leq x \leq 1 \\
 &= \frac{2}{\pi} \arcsin x.
 \end{aligned}$$

[Stirling's formula says that $n! \simeq e^{-n}n^{n+\frac{1}{2}}(2\pi)^{\frac{1}{2}}$ for large n .] This is an *arc-sine law*.

5.19 Example: Capture–Recapture

A population of b animals has had a number a of its members captured, marked, and released.

- (a) Let Y_m be the number of animals that it is necessary to capture (without re-release) to obtain m , which have been marked. Find $\mathbf{P}(Y_m = n)$ and $\mathbf{E}(Y_m)$.
- (b) If, instead, it had been decided just to capture $[\mathbf{E}(Y_m)]$ animals, what would have been the expected number of marked animals among them? Compare this with m .

Solution (a) **I** For the event $\{Y_m = n\}$ to occur, it is necessary that:

- (i) The n th animal is marked, which can occur in a ways.
- (ii) The preceding $n - 1$ animals include exactly $m - 1$ marked and $n - m$ unmarked animals, which may occur in $\binom{a-1}{m-1}\binom{b-a}{n-m}$ ways.

The total number of ways of first selecting a distinct animal to fill the n th place, and then choosing $n - 1$ animals to fill the remaining $n - 1$ places is $b \cdot \binom{b-1}{n-1}$. Because these are assumed to be equally likely, the required probability is

$$(1) \quad \mathbf{P}(Y_m = n) = \frac{a}{b} \binom{a-1}{m-1} \binom{b-a}{n-m} / \binom{b-1}{n-1}; \quad m \leq n \leq b - a + m.$$

To calculate $\mathbf{E}(Y_m)$, you may write

$$\begin{aligned} \mathbf{E}(Y_m) &= \sum_{n=m}^{b-a+m} n\mathbf{P}(Y_m = n) = \sum_{n=m}^{b-a+m} m \binom{a}{m} \binom{b-a}{n-m} / \binom{b}{n} \\ &= m \left(\frac{b+1}{a+1} \right)^{b+1-(a+1)+m} \sum_{n+1=m+1}^{b+1-(a+1)+m} \frac{a+1}{b+1} \binom{a}{m} \binom{b+1-(a+1)}{n+1-(m+1)} / \binom{b}{n} \\ &= m \left(\frac{b+1}{a+1} \right)^{b'-a'+m'} \sum_{n'=m'}^{b'-a'+m'} \frac{a'}{b'} \binom{a'}{m'} \binom{b'-a'}{n'-m'} / \binom{b'-1}{n'-1}, \end{aligned}$$

where $a' = a + 1$ and so on,

$$= m \left(\frac{b+1}{a+1} \right),$$

because (1) is a probability distribution with sum equal to unity.

II Alternatively, suppose that you were to capture them all, and let X_0 be the number of unmarked animals captured before the first marked animal, X_r the number of unmarked animals captured between the r th and the $(r + 1)$ st marked animals and X_a the number captured after the last marked animal. Then

$$\sum_0^a X_i = b - a,$$

and, by symmetry, for all i and j , $\mathbf{E}(X_i) = \mathbf{E}(X_j)$. Hence,

$$\mathbf{E}(X_r) = \frac{b-a}{a+1}, \quad \text{and} \quad \mathbf{E}(Y_m) = \sum_0^{m-1} \mathbf{E}(X_r) + m = m \frac{b+1}{a+1}.$$

(b) It is possible to write down the distribution of the number of marked animals captured, and then evaluate the mean by a method similar to the first method of (a). It is easier to let I_j be the indicator of the event that the j th captured animal is marked. Then the required expectation is

$$(2) \quad \mathbf{E} \left(\sum_1^{[\mathbf{E}(Y_m)]} I_j \right) = [\mathbf{E}(Y_m)] \mathbf{E}(I_j) = [\mathbf{E}(Y_m)] \frac{a}{b} = \left[\frac{b+1}{a+1} \cdot m \right] \frac{a}{b} < m.$$

Remark The distribution of Y_m is called the negative hypergeometric distribution, by analogy with the relation between the negative binomial distribution and the binomial distribution. The hypergeometric p.m.f. is (3.16.1).

- (3) **Exercise** If you capture and keep a fixed number n of animals, find the variance of the number that are marked.
- (4) **Exercise** Your pen will only hold m animals, so you return the unmarked ones. Now if Z_m is the number of captures required to secure m marked animals, find $\mathbf{E}(Z_m)$.
- (5) **Exercise** Let X and Y be independent binomial random variables with the same parameters n and p . Find $\mathbf{P}(X = k | X + Y = j)$ and explain why the answer takes the form you find.

5.20 Example: Visits of a Random Walk

Let $(S_n; n \geq 0)$ be a simple symmetric random walk with $S_0 = 0$.

- (a) Let V_r be the number of visits to r before the walk revisits the origin. Show that $\mathbf{E}(V_r) = 1$.
 (b) Show that the expected number of visits to the origin is infinite.

Solution Let I_n be the indicator of a visit to the point r at the n th step before any return to 0. Then

$$(1) \quad \begin{aligned} \mathbf{E}(V_r) &= \mathbf{E}\left(\sum_{n=1}^{\infty} I_n\right) = \sum_{n=1}^{\infty} \mathbf{E}(I_n) \\ &= \sum_{n=1}^{\infty} \mathbf{P}(S_n = r, S_1 \neq 0, \dots, S_{n-1} \neq 0). \end{aligned}$$

Now we make two important observations:

- (2) If $S_n = r$, then $X_1 + \dots + X_k \neq 0$ if and only if $X_{k+1} + \dots + X_n \neq r$.
 (3) Because the X_i are independent and identically distributed, $X_{k+1} + \dots + X_n$ has the same distribution as $X_1 + \dots + X_{n-k}$, and this remains true when $S_n = r$.
 Hence, we can write (1) as

$$(4) \quad \begin{aligned} \mathbf{E}(V_r) &= \sum_{n=1}^{\infty} \mathbf{P}\left(\sum_{i=1}^n X_i = r, X_1 \neq 0, \dots, X_1 + \dots + X_{n-1} \neq 0\right) \\ &= \sum_{n=1}^{\infty} \mathbf{P}\left(\sum_{i=1}^n X_i = r, X_2 + \dots + X_n \neq r, \dots, X_n \neq r\right) \quad \text{by (2),} \\ &= \sum_{n=1}^{\infty} \mathbf{P}\left(\sum_{i=1}^n X_i = r, X_1 + \dots + X_{n-1} \neq r, \dots, X_1 \neq r\right) \quad \text{by (3),} \\ &= \sum_{n=1}^{\infty} f_r(n) = 1, \quad \text{using Theorem 5.6.7.} \end{aligned}$$

- (b) Let J_n be the indicator of a visit to the origin at the n th step, and let R be the total number of returns to the origin. Then

$$\begin{aligned} \mathbf{E}(R) &= \mathbf{E}\left(\sum_{n=1}^{\infty} J_n\right) = \sum_{n=1}^{\infty} \mathbf{E}(J_n) = \sum_{n=1}^{\infty} \mathbf{P}(S_n = 0) \\ &= \sum_{k=1}^{\infty} \frac{1}{2^{2k}} \binom{2k}{k} \quad \text{where } n = 2k, \\ &= \sum_{k=1}^{\infty} \frac{(2k-1)(2k-3)\dots 3 \cdot 1}{2^k k(k-1)\dots 2 \cdot 1} \\ &\geq \sum_{k=1}^{\infty} (2k-1) \cdot \frac{(2k-2)}{(2k-1)} \cdot (2k-3) \cdot \frac{(2k-4)}{(2k-3)} \dots \frac{2}{3} \cdot \frac{1}{2^k k!} \\ &= \sum_{k=1}^{\infty} \frac{1}{2k} = \infty. \end{aligned}$$

Remark Result (a) is indeed remarkable. Interpreted as a game, it says that if a coin is tossed repeatedly and you get \$1 every time the total number of heads is r more than the total number of tails, until heads and tails are equal, then your expected gain is \$1, independently of r . [See Example 9.11 (b) for another method.]

- (5) **Exercise** For a symmetric simple random walk with $S_0 = 0$, let R_r be the total number of returns to r . What is $\mathbf{E}(R_r)$?
- (6) **Exercise** For a symmetric simple random walk with $S_0 = 0$, show that the probability that the first visit to S_{2n} takes place at time $2k$ is $\mathbf{P}(S_{2k} = 0)\mathbf{P}(S_{2n-2k} = 0); 0 \leq k \leq n$.
- (7) **Exercise** What is $\mathbf{E}(V)$, the expected number of visits of an asymmetric simple random walk to r ?
- (8) **Exercise** Consider a two-dimensional symmetric random walk (S_X, S_Y) on the points (i, j) , where i and j are integers. From (i, j) , the walk steps to any one of $(i \pm 1, j)$ or $(i, j \pm 1)$ with equal probability $\frac{1}{4}$. Show that the expected number $\mathbf{E}(V)$ of visits to the origin is infinite.

5.21 Example: Ordering

Let X and Y be random variables such that for all x

$$(1) \quad F_X(x) \leq F_Y(x).$$

Show that $\mathbf{E}(X) \geq \mathbf{E}(Y)$, and deduce that $F_X(x) \leq F_Y(x)$ if and only if, for all increasing functions $h(\cdot)$,

$$(2) \quad \mathbf{E}(h(X)) \geq \mathbf{E}(h(Y)).$$

Solution From Example 4.3.3, we have

$$(3) \quad \begin{aligned} \mathbf{E}(X) &= \sum_0^\infty \mathbf{P}(X > k) - \sum_{k=0}^{-\infty} \mathbf{P}(X < k) = \sum_0^\infty (1 - F_X(k)) - \sum_0^{-\infty} F_X(k) \\ &\geq \sum_0^\infty (1 - F_Y(k)) - \sum_0^{-\infty} F_Y(k) && \text{by (1)} \\ &= \mathbf{E}(Y). \end{aligned}$$

Now if $h(\cdot)$ is an increasing function

$$\begin{aligned} \mathbf{P}(h(X) > z) &= \mathbf{P}(X > \inf\{t: h(t) > z\}) \geq \mathbf{P}(Y > \inf\{t: h(t) > z\}) \quad \text{by (1)} \\ &= \mathbf{P}(h(Y) > z). \end{aligned}$$

Hence, $\mathbf{P}(h(X) \leq z) \leq \mathbf{P}(h(Y) \leq z)$ and (2) follows on using (3).

Conversely, if we choose $h(Z)$ to be the indicator of the event that $Z \leq x$, then

$$(4) \quad \mathbf{E}(h(X)) = \mathbf{P}(X \leq x) \leq \mathbf{P}(Y \leq x) = \mathbf{E}(h(Y)).$$

- (5) **Exercise** If X and Y are independent and for all x $F_X(x) \leq F_Y(x)$, show that $\mathbf{P}(X \geq Y) \geq \frac{1}{2}$.
- (6) **Exercise** If X, Y , and Z are independent show that X, Y , and Z can be distributed in such a way that $\mathbf{P}(X > Y) > \frac{1}{2}; \mathbf{P}(Y > Z) > \frac{1}{2}; \mathbf{P}(Z > X) > \frac{1}{2}$.
- (7) **Exercise** Let $X(n, p)$ have binomial distribution with parameters n and p . Show that

$$\mathbf{P}(X(m, p) \leq x) \geq \mathbf{P}(X(n, p) \leq x) \quad \text{for } m \leq n$$

and

$$\mathbf{P}(X(n, p_1) \leq x) \geq \mathbf{P}(X(n, p_2) \leq x) \quad \text{for } p_1 \leq p_2.$$

5.22 Example: More Martingales

Let $(X_n; n \geq 1)$ be a collection of independent random variables with respective means $(\mu_n; n \geq 1)$ and finite variances $(\sigma_n^2; n \geq 1)$. Show that

$$M_n = \left\{ \sum_{r=1}^n (X_r - \mu_r) \right\}^2 - \sum_{r=1}^n \sigma_r^2$$

defines a martingale with respect to X_n . Now assume that the X_n are identically distributed, with mean μ and variance σ^2 , and T is a stopping time for $(X_n; n \geq 1)$ with $\mathbf{E}T < \infty$. Show that when $Y_n = \sum_{r=1}^n X_r$,

$$\mathbf{E}(Y_T - T\mu)^2 = \sigma^2 \mathbf{E}T.$$

Solution First, by the independence,

$$\mathbf{E}|M_n| \leq \sum_{r=1}^n \mathbf{E}(X_r - \mu_r)^2 + \sum_{r=1}^n \sigma_r^2 < \infty.$$

Second, we have by using independence again,

$$\mathbf{E}(M_{n+1}|X_1, \dots, X_n) = M_n + \mathbf{E}(X_{n+1} - \mu_{n+1})^2 - \sigma_{n+1}^2 = M_n,$$

and M_n is a martingale.

For the last part, it is easy to see that we cannot apply the Optional Stopping theorem directly; so we employ an ingenious trick. First, note that $T \wedge n$ is a finite stopping time. By the first case in the optional stopping theorem 5.9.11, it follows that

$$(1) \quad \mathbf{E}(Y_{T \wedge n} - \mu T \wedge n)^2 = \sigma^2 \mathbf{E}T \wedge n.$$

Next, we observe that as $n \rightarrow \infty$, we have

$$(2) \quad T \wedge n \rightarrow T \quad \text{and} \quad Y_{T \wedge n} \rightarrow Y_T,$$

both statements being true with probability 1. Also, for any $m \geq n$, we have, using the fact that martingales have orthogonal increments, Corollary (5.9.14),

$$\begin{aligned} & \mathbf{E}(Y_{T \wedge m} - \mu T \wedge m - Y_{T \wedge n} + \mu T \wedge n)^2 \\ &= \mathbf{E}(Y_{T \wedge m} - \mu T \wedge m)^2 - \mathbf{E}(Y_{T \wedge n} - \mu T \wedge n)^2 \\ &= \sigma^2(\mathbf{E}T \wedge m - \mathbf{E}T \wedge n) \rightarrow 0 \quad \text{as } m, n \rightarrow \infty, \text{ since } \mathbf{E}T < \infty. \end{aligned}$$

Hence, by Theorem (5.7.4), $Y_{T \wedge n} - \mu n \wedge T$ converges in mean square as $n \rightarrow \infty$. But from (2) above, we know it converges to $Y_T - \mu T$ with probability 1. Hence,

$$\mathbf{E}(Y_{T \wedge n} - \mu T \wedge n)^2 \rightarrow \mathbf{E}(Y_T - \mu T)^2.$$

Now taking the limit as $n \rightarrow \infty$ in (1) gives the required result.

(3) **Exercise** Let $(X_n; n \geq 1)$ be independent with respective finite means $(\mu_n; n \geq 1)$. Show that U_n and V_n are martingales, where

$$(a) U_n = \sum_{r=1}^n X_r - \sum_{r=1}^n \mu_r. \quad (b) V_n = \frac{X_1 \cdots X_n}{\mu_1 \cdots \mu_n}.$$

(4) **Exercise** Let $D_n = X_n - X_{n-1}$, where X_n is a martingale with finite variance. Show that

$$\text{var } X_n = \sum_{r=1}^n \text{var } D_r.$$

(5) **Exercise** Let $(X_n; n \geq 1)$ and $(Y_n; n \geq 1)$ be two collections of independent random variables, with each collection identically distributed having respective means μ_x and μ_y . Show that if T is a stopping time with respect to the sequence $\{(X_n, Y_n); n \geq 1\}$ and $\mathbf{E}T < \infty$, then

$$\mathbf{E} \left[\left(\sum_1^T X_r - T\mu_x \right) \left(\sum_1^T Y_r - T\mu_y \right) \right] = \mathbf{E}T \text{cov}(X_1, Y_1).$$

(6) **Exercise** Let $(X_n; n \geq 1)$ be independent and identically distributed with $M(t) = \mathbf{E}e^{tX_1} < \infty$. Set $S_n = \sum_1^n X_r$. Show that $M_n = \exp(tS_n)(M(t))^{-n}$ is a martingale with respect to S_n .

5.23 Example: Simple Random Walk Martingales

Let $S_n = S_0 + X_1 + \cdots + X_n$, where $(X_n; n \geq 1)$ are independent, and such that

$$0 \neq \mathbf{P}(X_1 = 1) = p = 1 - q = 1 - \mathbf{P}(X_1 = -1) \neq 1$$

- (a) Show that $(q/p)^{S_n}$ is a martingale with respect to S_n .
- (b) If $a < S_0 < b$, find the probability that the walk hits a before it hits b , where a, b and S_0 are integers.

Solution (a) We noted in (5.6.10) that $\mathbf{E}(\frac{q}{p})^{X_r} = 1$. Because the X_n are independent, the conditions for $(\frac{q}{p})^{S_n}$ to be a martingale follow easily.

Let T be the first time at which S_n takes either of the values a or b . The probability that any consecutive sequence of X_n , of length $a + b$, yields $a + b$ consecutive 1s is p^{a+b} . Hence,

$$\mathbf{P}(T > m(a + b)) < (1 - p^{a+b})^m \rightarrow 0, \quad \text{as } m \rightarrow \infty.$$

Hence, $\mathbf{P}(T < \infty) = 1$. Because S_n and $S_{n \wedge T}$ are bounded, so are $(\frac{q}{p})^{S_n}$ and $(\frac{q}{p})^{S_{n \wedge T}}$. Let A be the event that the walk hits a before b ; because $\mathbf{P}(T < \infty) = 1$, we must have

$$\mathbf{P}(\text{walk hits } b \text{ before } a) = 1 - \mathbf{P}(A).$$

We can apply the second case of the optional stopping theorem 5.9.11 to obtain, if $S_0 = s$,

$$\begin{aligned} \left(\frac{q}{p}\right)^s &= \mathbf{E}\left(\frac{q}{p}\right)^{S_0} = \mathbf{E}\left(\frac{q}{p}\right)^{S_T} \\ &= \left(\frac{q}{p}\right)^a \mathbf{P}(A) + \left(\frac{q}{p}\right)^b (1 - \mathbf{P}(A)). \end{aligned}$$

Hence,

$$\mathbf{P}(A) = \frac{\left(\frac{q}{p}\right)^s - \left(\frac{q}{p}\right)^b}{\left(\frac{q}{p}\right)^a - \left(\frac{q}{p}\right)^b}.$$

- (1) **Exercise** If $p = q = \frac{1}{2}$, show that S_n is a martingale and deduce that

$$\mathbf{P}(A) = \frac{s-b}{a-b}.$$

- (2) **Exercise** If $p = q = \frac{1}{2}$, show that $S_n^2 - n$ is a martingale and hence that

$$\mathbf{E}T = (s-a)(b-s).$$

- (3) **Exercise** If $a \rightarrow -\infty$ and $p > \frac{1}{2}$, use an appropriate martingale to show that

$$\mathbf{E}T = \frac{(b-s)}{1-2q}.$$

- (4) **Exercise** Show that $[(S_n - n(p-q))^2 - 4npq; n \geq 0]$ is a martingale. If $a \rightarrow -\infty$ and $p > \frac{1}{2}$, deduce that

$$\text{var}T = \frac{4(b-s)pq}{(p-q)^3}.$$

- (5) **Exercise** Let S_n be a simple symmetric random walk started at the origin, and let T be the number of steps until the walk first hits $-a$ or b , where a and b are positive. Show that the following are all martingales:

(a) S_n ; (b) $S_n^2 - n$; (c) $S_n^3 - 3nS_n$; (d) $S_n^4 - 6nS_n^2 + 3n^2 + 2n$.

Hence find $\mathbf{P}(S_T = -a)$, $\mathbf{E}T$, and $\mathbf{E}(T \cap \{S_T = -a\})$. Show finally that $\text{var}T = ab(a^2 + b^2 - 2)/3$.

5.24 Example: You Can't Beat the Odds

Let Y_n be the total net fortune of a gambler after betting a unit stake on each of n consecutive fair plays in a casino. Thus, the return from the n th unit stake is $Y_n - Y_{n-1}$. We assume that Y_n constitutes a martingale. A gambler devises a betting system that entails placing a stake S_n on the n th play, where S_n is not necessarily a unit stake, but S_n is necessarily a function only of (Y_0, \dots, Y_{n-1}) , and does not depend on any Y_{n+k} , $k \geq 0$. Write down an expression for the gambler's fortune Z_n , after n plays, and show that Z_n is a martingale if $\mathbf{E}|Z_n| < \infty$.

Solution From the description of the system, the return on the n th play is $S_n(Y_n - Y_{n-1})$. Hence,

$$Z_n = Z_{n-1} + S_n(Y_n - Y_{n-1}) = Y_0 + \sum_{r=1}^n S_r(Y_r - Y_{r-1}).$$

Because S_n is a function of (Y_0, \dots, Y_{n-1}) , we have

$$\begin{aligned} \mathbf{E}(Z_{n+1}|Y_0, \dots, Y_n) &= Z_n + \mathbf{E}(S_{n+1}(Y_{n+1} - Y_n)|Y_0, \dots, Y_n) \\ &= Z_n + S_{n+1}[\mathbf{E}(Y_{n+1}|Y_0, \dots, Y_n) - Y_n] \\ &= Z_n \end{aligned}$$

because Y_n is a martingale. The result follows.

Remark The point of this example is that you cannot turn a fair game in your favour using a system. The exercises supply more instances of this.

- (1) **Exercise** Show that using any of the following systems, the gambler's fortune is a martingale:
 - (a) *Optional skipping.* At each play, the gambler skips the round or wagers a unit stake.
 - (b) *Optional starting.* The gambler does not join in until the $(T + 1)$ th play, where T is a stopping time for Y_n .
 - (c) *Optional stopping.* The gambler uses the system until a stopping time T , and then quits.
- (2) **Exercise:** *Optional Sampling* The gambler only uses the system at the plays numbered (T_1, T_2, \dots) , where $(T_n; n \geq 1)$ is a sequence of stopping times such that $\mathbf{P}(T_r \leq n_r) = 1$ for some non random sequence of finite real numbers n_r . Show that $(Z_{T_r}; r \geq 1)$, where $T_0 = 0$, is a martingale.
- (3) **Exercise** Show that the result of Exercise (2) is true if the gambler's fortunes are bounded by $K < \infty$, and she plays only at stopping times $(T_r; r \geq 1)$, where

$$0 \leq T_1 \leq T_2 \leq T_3 \leq \dots$$

5.25 Example: Matching Martingales

In a cloakroom, there are C distinct coats belonging to C people who all attempt to leave by picking a coat at random. Those who select their own coat leave; the rest return their coats and pick again at random. This continues until everyone leaves; let N be the number of rounds required. Show that $\mathbf{E}N = C$ and $\text{var}N \leq C$.

Solution Let M_n be the number of people present after the n th round, and X_n the number of matches in the n th round. Thus, $M_0 = C$, $M_{n+1} = M_n - X_{n+1}$, $n \geq 0$, and $M_N = 0$. By the result of Example 5.4.3, $\mathbf{E}X_n = 1$ for all n , so that

$$\mathbf{E}(M_{n+1} + n + 1|M_0, \dots, M_n) = M_n + n.$$

Thus, $(M_n + n; n \geq 0)$ is a martingale, and N is clearly a stopping time. Also, $\mathbf{P}(\text{at least one match}) \geq C^{-1}$ for all values of M_n , so $\mathbf{P}(N < \infty) = 1$, and also $\mathbf{E}N < \infty$. By the appropriate part of the optional stopping theorem,

$$C = M_0 + 0 = \mathbf{E}(M_N + N) = \mathbf{E}N.$$

We also have from Example 5.4.3 that

$$\text{var}(X_{n+1}|M_0, \dots, M_n) = \begin{cases} 1 & \text{if } M_n > 1 \\ 0 & \text{if } M_n = 1 \end{cases}.$$

Hence, $\text{var } X_{n+1} \leq 1$, and we may write

$$\begin{aligned} & \mathbf{E}((M_{n+1} + n + 1)^2 + M_{n+1} | M_0, \dots, M_n) \\ &= (M_n + n)^2 - 2(M_n + n)\mathbf{E}(X_{n+1} - 1) + M_n \\ & \quad + \mathbf{E}((X_{n+1} - 1)^2 - X_{n+1} | M_0, \dots, M_n) \\ & \leq (M_n + n)^2 + M_n. \end{aligned}$$

Thus, $(M_n + n)^2 + M_n$ is a nonnegative supermartingale, and by the appropriate optional stopping theorem [given at the end of Section 5.7],

$$C^2 + C = M_0^2 + M_0 \geq \mathbf{E}((M_N + N)^2 + M_N) = \mathbf{E}N^2.$$

The result follows, using the first part.

- (1) **Exercise** Suppose the coat-grabbers adopt a slightly smarter approach. At each round, those with their own coats leave, those left call out the name on the label of the coat they have picked. Any pair holding each other's coat swap them, and leave. The rest return their coats for another round.
- (a) Show that the expected number of rounds now required is $C/2$.
- (b) Let X_n be the number of departures in the n th round; show that $\text{var} X_n = 3$, for $M_n \geq 4$. [Hint: With an obvious notation using suitable indicators, $X_n = \sum_j I_j + \sum_{j \neq k} I_{jk}$. Hence, when $M_{n-1} = m$,

$$\begin{aligned} \mathbf{E}X_n^2 &= m\mathbf{E}I_1 + m(m-1)\mathbf{E}(I_1 I_2) \\ & \quad + 2m(m-1)(m-2)\mathbf{E}(I_1 I_{23}) + 2m(m-1)\mathbf{E}I_{12} \\ & \quad + m(m-1)(m-2)(m-3)\mathbf{E}(I_{12} I_{34}) = 7. \end{aligned}$$

Show that $(M_n + 2n)^2 + \frac{3}{2}M_n$ is a supermartingale, and deduce that $\text{var} N \leq \frac{3}{2}C$.

- (2) **Exercise** Suppose it was a mathematicians party, and at each round any subgroup of size less than or equal to k , that holds no coats outside the subgroup, simply redistributes their coats correctly, and leaves. Show that the expected number of rounds required is C/k .
- (3) **Exercise** Suppose now that the purpose of the coats exercise is not simply to leave, but to leave in pairs. Thus, only pairs holding each others coat swap and leave; the rest, including those who have their own coat, return them for another round. Show that when C is even, $\mathbf{E}N = C$ and $\text{var} N \leq 2C$. What can you say when C is odd?

5.26 Example: Three-Handed Gambler's Ruin

Three players start with a , b , and c chips, respectively, and play the following game. At each stage, two players are picked at random, and one of those two is picked at random to give the other a chip. This continues until one of the three is out of chips, and quits the game; the other two continue until one player has all the chips. Let X_n , Y_n , and Z_n be the chips possessed by each player, respectively, after the n th stage; and let T be the number of transfers until someone has all the chips. Show that

$$\mathbf{E}T = ab + bc + ca.$$

Solution We claim that

$$U_n = X_n Y_n + Y_n Z_n + Z_n X_n + n$$

is a martingale. To see this, we need to consider two cases; thus,

(i) $X_n Y_n Z_n > 0$. Here, for instance,

$$\begin{aligned} \mathbf{E}(X_{n+1} Y_{n+1} | X_0, Y_0, Z_0, \dots, X_n, Y_n, Z_n) \\ &= \frac{1}{6} [(X_n + 1)Y_n + (X_n - 1)Y_n + X_n(Y_n + 1) + X_n(Y_n - 1) \\ &\quad + (X_n + 1)(Y_n - 1) + (Y_n + 1)(X_n - 1)] \\ &= X_n Y_n - \frac{1}{3}. \end{aligned}$$

The other two terms being treated similarly, we find that

$$(1) \quad \mathbf{E}(U_{n+1} + n + 1 | X_0, \dots, Z_n) = U_n + n + 1 - 1.$$

(ii) One of X_n, Y_n or Z_n is zero. If for instance $Z_n = 0$, then

$$\mathbf{E}(X_{n+1} Y_{n+1} | X_0, \dots, Z_n) = \frac{1}{2} [(X_n + 1)(Y_n - 1) + (X_n - 1)(Y_n + 1)] = X_n Y_n - 1.$$

The other two possibilities being treated similarly, we obtain the same martingale condition (1). Clearly, T is a finite-mean stopping time for this bounded martingale, and $U_T = 0$, so

$$\mathbf{E}T = \mathbf{E}(U_T + T) = \mathbf{E}(U_0 + 0) = ab + bc + ca.$$

(2) **Exercise** Let S be the number of transfers until one of the players is first out of chips. Show that

$$M_n = X_n Y_n Z_n + \frac{1}{3}n(a + b + c)$$

is a martingale, and deduce that

$$\mathbf{E}S = \frac{3abc}{a + b + c}.$$

(3) **Exercise** The three players play a different game. Thus, they start with a, b , and c chips, respectively. At each stage, one player is picked at random to receive one chip from each other player still in; players drop out when they have no chips. Show that M_n and V_n are martingales, where

$$\begin{aligned} M_n &= X_n Y_n Z_n + n(a + b + c - 2) \\ V_n &= X_n Y_n + Y_n Z_n + Z_n X_n + 3n. \end{aligned}$$

If S and T are defined as above, deduce that

$$\mathbf{E}S = \frac{abc}{a + b + c - 2},$$

and

$$\mathbf{E}T = ab + bc + ca - \frac{2abc}{a + b + c - 2}.$$

(4) **Exercise** The three players are now joined by a fourth, and all four return to play by the rules of the first game. The fourth starts with d chips, and they have X_n, Y_n, Z_n , and W_n at the n th stage. Let S be the first time at which only two players remain in the game, and T the first time at which only one is left with all the chips. Verify that

$$U_n = X_n Y_n + Y_n Z_n + Z_n X_n + X_n W_n + W_n Y_n + W_n Z_n + n,$$

and

$$V_n = X_n Y_n Z_n + W_n X_n Y_n + W_n Y_n Z_n + W_n X_n Z_n + \frac{n}{2}(a + b + c + d)$$

are martingales, and deduce that

$$\mathbf{E}S = \frac{2(abc + bcd + acd + abd)}{a + b + c + d},$$

and

$$\mathbf{E}T = ab + bc + cd + da + ac + bd.$$

PROBLEMS

- 1 You roll two fair dice. Let X be the number of 2s shown, and Y the number of 4s. Write down the joint probability mass function of X and Y , and find $\text{cov}(X, Y)$ and $\rho(X, Y)$.
- 2 Let the random variables X and Y have joint probability mass function $f(x, y)$ such that:

$$\begin{aligned} f(1, 2) &= \frac{1}{8}, & f(1, 3) &= \frac{1}{16}, & f(1, 4) &= \frac{1}{4}, \\ f(2, 2) &= \frac{1}{16}, & f(2, 3) &= \frac{1}{8}, & f(2, 4) &= \frac{3}{8}. \end{aligned}$$

Find the probability of the following:

- (a) $X > Y$
 - (b) $X \geq Y$
 - (c) $X + Y$ is odd
 - (d) $X - Y \leq 1$.
- 3 Find two random variables X and Y that are uncorrelated, but not independent.
 - 4 Show that if $\mathbf{E}((X - Y)^2) = 0$, then $X = Y$ with probability one.
 - 5 Show that if $\mathbf{E}((X - Y)^2) = \mathbf{E}(X^2) + \mathbf{E}(Y^2)$, then X and Y are orthogonal.
 - 6 Let X be uniformly distributed on $\{0, 1, 2, \dots, 4n\}$. Let $Y = \sin(\frac{1}{2}\pi X)$ and $Z = \cos(\frac{1}{2}\pi X)$.
 - (a) What is the joint probability mass function of Y and Z ?
 - (b) What is the distribution of $Y + Z$?
 Show that Y and Z are orthogonal.
 - 7 Let X and Y be jointly distributed with finite second moments and unit variance. Show that for some nonzero constants a, b, c, d , the random variables U and V are uncorrelated where $U = aX + bY$, $V = cX + dY$. Are a, b, c , and d unique?
 - 8 A source produces a message forming a sequence of zeros and ones. In being transmitted, it passes through two independent channels, each of which transmits the wrong symbol with probability $1 - p$, or the correct symbol with probability p . Show that a symbol is least likely to be transmitted correctly when $p = \frac{1}{2}$.

Find the probability of correct transmission of a symbol when the message passes through three similar independent channels.
 - 9 Let $(X_n; n \geq 1)$ be a sequence of independent random variables such that $\mathbf{P}(X_n = 1) = p = 1 - q = 1 - \mathbf{P}(X_n = -1)$. Let U be the number of terms in the sequence before the first change of sign, and V the further number of terms before the second change of sign. (In other words, X_1, X_2, \dots

is made up of runs of +1s and runs of -1s; U is the length of the first run and V the length of the second.)

(a) Show that $\mathbf{E}(U) = pq^{-1} + qp^{-1}$ and $\mathbf{E}(V) = 2$.

(b) Write down the joint distribution of U and V , and find $\text{cov}(U, V)$ and $\rho(U, V)$.

- 10** An urn contains n balls numbered individually with the integers from 1 to n . Two balls are drawn at random without replacement, and the numbers they bear are denoted by X and Y . Find $\text{cov}(X, Y)$, $\rho(X, Y)$, and the limit of $\rho(X, Y)$ as $n \rightarrow \infty$.

- 11** Let X and Y have joint distribution defined by $f(0, 0) = 1 - 3a$; and $f(0, 1) = f(1, 0) = f(1, 1) = a$; $a \leq \frac{1}{3}$. Find:

(a) The p.m.f.s of X and Y

(b) $\text{cov}(X, Y)$

(c) $\mathbf{E}(X|Y)$ and $\mathbf{E}(Y|X)$

(d) Whether X and Y can be independent, and if so, when.

- 12** You roll two fair dice, and they show X and Y , respectively. Let $U = \min\{X, Y\}$, $V = \max\{X, Y\}$. Write down the joint distributions of:

(a) $\{U, X\}$

(b) $\{U, V\}$

(c) $\{X, Y, V\}$.

Find $\text{cov}(U, V)$ and $\mathbf{E}(XYV)$.

- 13** (a) If X and Y are independent with finite expectation, show that $\mathbf{E}(XY)$ exists.

(b) Find a sufficient condition on the moments of X and Y , for $\mathbf{E}(XY)$ to exist in general.

- 14** Which of the following functions $f(i, j)$ can be a joint probability mass function of two random variables X and Y ?

(a) $\theta^{|i|+|j|}$;

$|i| + |j| \neq 0$

(b) θ^{i+j} ;

$0 \leq i \leq j < \infty$

(c) θ^{i+j+2} ;

$0 \leq i < j < \infty$

(d) θ^{i+j+1} ;

$0 \leq i, j < \infty$

(e) $(i^j - (i-1)^j)\alpha \left(\frac{\beta}{c}\right)^j$;

$1 \leq i \leq c, j \geq 1, c$ an integer

(f) $\alpha(i^n - (i-1)^n)j^{-n-2}$;

$1 \leq i \leq j < \infty$.

Are X and Y independent in any case?

- 15** For each function in problem 14 that is a joint probability mass function of X and Y , find the marginal mass functions of X and Y .

- 16** Suppose that random variables X and Y are such that $\mathbf{P}(|X - Y| \leq M) = 1$, where M is finite. Show that if $\mathbf{E}(X) < \infty$, then $\mathbf{E}(Y) < \infty$ and $|\mathbf{E}(X) - \mathbf{E}(Y)| \leq M$.

- 17** Show that the following are joint p.m.f.s and find the marginal distributions.

(a)

$$f(x_1, \dots, x_k) = \frac{x!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}, \text{ where } \sum_1^k p_j = 1 \text{ and } \sum_1^k x_j = x.$$

(b)

$$f(x, x_1, \dots, x_k) = \frac{(x+r-1)!}{(r-1)!} p_0^r \frac{p_1^{x_1}}{x_1!} \dots \frac{p_k^{x_k}}{x_k!},$$

where $\sum_0^k p_j = 1$ and $\sum_1^k x_j = x \geq 0$.

(c)

$$f(x_1, \dots, x_k) = \binom{a_1}{x_1} \cdots \binom{a_k}{x_k} / \binom{\sum_{i=1}^k a_i}{\sum_{i=1}^k x_i}.$$

- 18 Let X and Y be independent geometric random variables with parameters p_1 and p_2 , respectively.
 (a) If c is an integer and $Z = \min\{c, X\}$, find $\mathbf{E}(Z)$.
 (b) Find the distribution and expectation of $\min\{X, Y\}$.
- 19 Let X and Y have joint p.m.f.

$$f(x, y) = \frac{C}{(x + y - 1)(x + y)(x + y + 1)}; \quad m \geq 1, n \geq 1.$$

Find the p.m.f. of X , the p.m.f. of Y , and the value of C .

- 20 Let X and Y be independent random variables each with a geometric distribution, so

$$f_X(x) = \alpha\beta^{x-1}; \quad x \geq 1, \alpha + \beta = 1,$$

$$f_Y(y) = pq^{y-1}; \quad y \geq 1, p + q = 1.$$

Let $R = X/Y$.

- (a) Find $\mathbf{P}(R > 1)$.
 (b) If $r = m/n$ where m and n are integers with no common factor except unity, find $\mathbf{P}(R = r)$, and show that when $\alpha = p = \frac{1}{2}$, $\mathbf{P}(R = r) = 1/(2^{m+n} - 1)$.
- 21 **Triangular Distribution** Let X and Y be independent random variables each uniformly distributed on $\{0, 1, \dots, n\}$. Find the p.m.f. of
 (a) $X + Y$.
 (b) $X - Y$.
- 22 Let X have binomial distribution with parameters n and p , and Y a binomial distribution with parameters m and p . Show that if X and Y are independent then $X + Y$ has a binomial distribution with parameters $m + n$ and p . Deduce that

$$\sum_{k=0}^r \binom{m}{r-k} \binom{n}{k} = \binom{m+n}{r}.$$

Find the conditional distribution of X given that $X + Y = k$.

- 23 If X has a binomial distribution with parameters n and p , show that

$$\mathbf{E}\left(\frac{1}{1+X}\right) = \frac{1 - (1-p)^{n+1}}{(n+1)p}.$$

- 24 Let X and Y be independent Poisson random variables. Show that $Z = X + Y$ has a Poisson distribution.

Show also that for some p , $\mathbf{P}(X = k|Z = n) = \binom{n}{k} p^k (1-p)^{n-k}$, which is to say that the conditional distribution of X given Z is binomial.

- 25 Let X and Y be independent geometric random variables, such that for $m \geq 0$

$$\mathbf{P}(X = m) = (1 - \lambda)\lambda^m \quad \text{and} \quad \mathbf{P}(Y = m) = (1 - \mu)\mu^m.$$

(a) Show that

$$\mathbf{P}(X + Y = n) = \frac{(1 - \lambda)(1 - \mu)}{\lambda - \mu} (\lambda^{n+1} - \mu^{n+1}), \quad \lambda \neq \mu.$$

Find $\mathbf{P}(X = k|X + Y = n)$.

(b) Find the distribution of $Z = X + Y$ when $\lambda = \mu$, and show that in this case $\mathbf{P}(X = k|Z = n) = 1/(n + 1)$.

26 Bell's Inequality Let X, Y , and Z be jointly distributed random variables such that each can take either of the values ± 1 . Show that $\mathbf{E}(XY) \leq 1 - |\mathbf{E}((X - Y)Z)|$.

(This inequality is interesting because it has been claimed that there are experiments in quantum mechanics for which it does not hold true.)

27 Show that for any c such that $|c| \leq 4$ the function

$$f(i, j) = \frac{1}{(2m + 1)(2n + 1)} + \frac{c(i - m)(j - n)}{((2n + 1)(2m + 1))^2}; \quad \begin{array}{l} 0 \leq i \leq 2m, \\ 0 \leq j \leq 2n, \end{array}$$

is a joint probability mass function with marginal distributions that do not depend on c .

Show that the covariance of this distribution is

$$\frac{cmn(m + 1)(n + 1)}{9(2n + 1)(2m + 1)}.$$

28 Construct two identically distributed random variables X and Y such that

$$\mathbf{P}(X < Y) \neq \mathbf{P}(Y < X).$$

29 Bernoulli's Urn Initially an urn contains U umber balls and a vase contains V viridian balls. From each container, a ball is removed at random and placed in the other container. Let U_r be the number of umber balls in the urn after r repetitions of this operation.

(a) Find $\mathbf{E}(U_r)$ and show that $\lim_{r \rightarrow \infty} \mathbf{E}(U_r) = U^2/(U + V)$.

(b) Just before each time balls are exchanged, a coin is tossed (which shows a head with probability p); find the expected number of umber balls in the urn when the coin first shows a head. Show that if $U = V$ and $Up = 1$, this expectation is about $\frac{2}{3}U$ for large U .

30 Let $(X_i; i \geq 1)$ be a random walk with $S_0 = 0$ and $S_r = \sum_1^r X_i$. Define the maximum $M_n = \max_{1 \leq k \leq n} \{S_k\}$. Show that $\mathbf{P}(M_n \geq x) \leq (\frac{p}{q})^x$, $x \geq 0$, and deduce that

$$\lim_{n \rightarrow \infty} \mathbf{E}(M_n) \leq q/(q - p), \quad \text{for } p < q.$$

31 An urn contains n balls such that each of the n consecutive integers $1, 2, \dots, n$ is carried by one ball. If k balls are removed at random, find the mean and variance of the total of their number in the two cases:

(a) They are not replaced.

(b) They are replaced.

What is the distribution of the largest number removed in each case?

32 Let the random variables X and Y have joint distribution

$$\begin{aligned} \mathbf{P}(X = a, Y = 0) &= \mathbf{P}(X = 0, Y = a) = \mathbf{P}(X = -a, Y = 0) \\ &= \mathbf{P}(X = 0, Y = -a) = \frac{1}{4}. \end{aligned}$$

Show that $X - Y$ and $X + Y$ are independent.

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

$$\begin{aligned} \mathbf{P}(U = +1) &= \mathbf{P}(U = -1) = \frac{1}{2}, \\ \mathbf{P}(V = +1|U = 1) &= \frac{1}{3} = \mathbf{P}(V = -1|U = -1), \\ \mathbf{P}(V = -1|U = 1) &= \frac{2}{3} = \mathbf{P}(V = +1|U = -1). \end{aligned}$$

- (a) Find the probability that $x^2 + Ux + V = 0$ has at least one real root.
 (b) Find the expected value of the larger root given that there is at least one real root.
 (c) Find the probability that $x^2 + (U + V)x + U + V = 0$ has at least one real root.
- 34 Let $S_n = \sum_1^n X_i$ be a random walk with $S_0 = a > 0$, such that $\mathbf{P}(X_i = -1) = q$, $\mathbf{P}(X_i = +2) = p$, $p + q = 1$. Let T_{a0} be the time at which the walk first visits zero. Show that if $p \leq \frac{1}{3}$ then $\mathbf{P}(T_{a0} < \infty) = 1$, but if $p > \frac{1}{3}$ then $\mathbf{P}(T_{a0} < \infty) = r^a < 1$. What is r ?

- 35 Casualties arriving at a certain hospital require surgery, independently of one another, with probability $\frac{1}{4}$. What is the probability that, on a day when n casualties arrive, exactly r require surgery?

The number X of casualties arriving on weekdays follows a Poisson distribution with mean 8; that is, for each day, $\mathbf{P}\{X = n\} = e^{-8} 8^n / n!$, $n = 0, 1, 2, \dots$

Show that the number requiring surgery each day also follows a Poisson distribution and find its mean.

Suppose that the situation is identical on Saturdays and Sundays, except that there are on average only four casualties arriving per day. Find the mean and variance of the number of patients requiring surgery each week. (Assume that each day's arrivals are independent and recall Problem 24.)

- 36 An urn contains m white balls and $M - m$ black balls. Balls are chosen at random without replacement. Show that the probability p_k of choosing exactly k white balls in n choices ($0 \leq k \leq m$) is given by

$$p_k = \binom{M}{n}^{-1} \binom{m}{k} \binom{M-m}{n-k}.$$

Define a random variable $X = X_1 + X_2 + \dots + X_n$, where $X_i = 0$ or 1 according as the i th ball is black or white. Show that

$$\begin{aligned} \mathbf{P}(X = k) &= p_k, \\ \mathbf{P}(X_i = 1) &= m/M, \\ \mathbf{P}(X_i = 1, X_j = 1) &= \frac{m(m-1)}{M(M-1)}, \quad i \neq j. \end{aligned}$$

By considering $\mathbf{E}(X)$, $\mathbf{E}(X^2)$, or otherwise, find the mean and variance of the distribution given by p_k .

- 37 **Conditional Gambler's Ruin** An optimistic gambler seeks to know the expected duration of the game assuming that he wins. As usual, he plays a sequence of fair wagers losing or gaining \$1 each time. The game stops as soon as he has \$0 or \$ K . Initially, his fortune is \$ k ($k < K$), the event that he stops with \$ K is V_k , and D_k is the duration of the game. Let $\delta_k = \mathbf{E}(D_k | V_k)$. Show that for $1 < k < K$,

$$(k+1)\delta_{k+1} - 2k\delta_k + (k-1)\delta_{k-1} + 2k = 0.$$

Write down two boundary conditions at $k = 1$ and $k = K$, and deduce that

$$\mathbf{E}(D_k | V_k) = \frac{1}{3}(K^2 - k^2), \quad 1 \leq k \leq K.$$

- 38 Let $(S_n; n \geq 1)$ be a simple random walk, and let M be its maximum, $M = \max_{n \geq 1} \{S_n\}$.
 (a) If $S_0 = 0$, and $p < q$, show that m has a geometric distribution and find its mean.
 (b) If S_0 is a random variable with distribution $\mathbf{P}(S_0 = -k) = \alpha\beta^k$; $k = 0, 1, 2, \dots$ find the distribution of M .

In this case, what is the conditional distribution of S_0 given M ?

- 39 Let X_1, X_2 , and X_3 be independent geometric random variables with parameters $1 - p_1, 1 - p_2$, and $1 - p_3$, respectively.

(a) Show that

$$\mathbf{P}(X_1 < X_2 < X_3) = \frac{(1 - p_1)(1 - p_2)p_2p_3^2}{(1 - p_2p_3)(1 - p_1p_2p_3)}.$$

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

(c) Three players, A , B , and C , roll a fair die in turn, that is, in the order $ABCABCA \dots$. Show that the probability that A throws the first six, B the second six, and C the third six, is $\frac{216}{1001}$.

40 Matching Once again, n letters with n matching envelopes are inadvertently placed at random in the envelopes. Let X be the number of letters that are in their matching envelope. Find $\mathbf{E}(X)$ and $\text{var}(X)$, and show that

$$\mathbf{E}(X(X-1)\dots(X-k+1)) = \begin{cases} 1 & k \leq n \\ 0 & k > n. \end{cases}$$

41 Let n be a prime number greater than two, and let X and Y be independently and uniformly distributed on $\{0, 1, \dots, n-1\}$. For all r such that $0 \leq r \leq n-1$, define $Z_r = X + rY$, modulo n . Show that the random variables $(Z_r; 0 \leq r \leq n-1)$ are pairwise independent.

Is this true if n is not prime?

42 Jensen's Inequality If $g(\cdot)$ is convex, show that $g(\mathbf{E}(X|Y)) \leq \mathbf{E}(g(X)|Y)$.

43 Polya's urn (Example 2.7) revisited. A bag initially contains r red and b blue balls, $rb > 0$. A ball is drawn at random, its colour noted, and it is returned to the bag together with a new ball of the same colour. Let $R(n)$ be the number of red balls in the bag after n such operations. Let T be the number of balls drawn until the first blue ball appears.

(a) Show that $R(n)/(n+b+r)$ is a martingale.

(b) Deduce that $\mathbf{E}\{(b+1)(b+r)/(T+r+b)\} = b$.

44 Optional Stopping. Let $X(n)$ be a martingale, and T a stopping time for $X(n)$ such that $\mathbf{P}(T < \infty) = 1$. Prove that $\mathbf{E}X(T) = \mathbf{E}X(0)$ if either of (a) or (b) holds.

(a) $\mathbf{E}(\sup_n |X(n \wedge T)|) < \infty$.

(b) $\mathbf{E}|X(T)| < \infty$, and $\mathbf{E}\{X(n)I(T > n)\} \rightarrow 0$ as $n \rightarrow \infty$.

6

Generating Functions and Their Applications

Everything future is to be estimated by a wise man, in proportion to the probability of attaining it, and its value when attained.

Samuel Johnson, [The Rambler, 20]

This chapter deals with a special subject and may be omitted on a first reading. Its contents are important and useful, but are not a prerequisite for most of the following chapters.

6.1 Introduction

In Chapter 3, we found that generating functions can provide elegant and concise methods for handling collections of real numbers. The mass function of an integer valued random variable is such a collection, and so we may anticipate (correctly as it turns out) that the following generating function will be useful.

- (1) **Definition** *The probability generating function $G(s)$ of the integer valued random variable X is defined by*

$$G(s) = \sum_k \mathbf{P}(X = k)s^k. \quad \blacktriangle$$

Because all random variables in this chapter are integer valued, this is not again mentioned explicitly.

- (2) **Example** Let X be uniformly distributed in $\{-a, -a + 1, \dots, b - 1, b\}$, where $a, b > 0$. Then provided $s \neq 1$,

$$G(s) = \sum_{k=-a}^b \frac{1}{a+b+1} s^k = \frac{s^{-a} - s^{b+1}}{(a+b+1)(1-s)}. \quad \bullet$$

Notice that, by Theorem 4.3.4, we have from Definition 1 of $G(s)$ that

- (3)
$$G(s) = \mathbf{E}(s^X);$$

this is a particularly useful representation of $G(s)$, and we use it a great deal in what follows.

For example, suppose we seek the probability generating function of $Y = X + a$, where a is constant. Using (3), we can write

$$G_Y(s) = \mathbf{E}(s^Y) = \mathbf{E}(s^{X+a}) = s^a G_X(s).$$

We will see many other applications of (3) later. When X is defective (that is, when $\mathbf{P}(|X| < \infty) < 1$), the representation (3) can still be used, provided that we remember that the expectation is taken over the finite part of the distribution of X .

We write $G_X(s)$ when we want to stress the role of X ; and for brevity, $G_X(s)$ is sometimes known as the p.g.f. of X .

Obviously, if $\mathbf{P}(|X| < \infty) = 1$, then

$$G_X(1) = \sum_k \mathbf{P}(X = k) = 1.$$

To sum up, if X is finite with probability 1, then $G_X(s)$ is a power series in s with nonnegative coefficients such that $G_X(1) = 1$.

Conversely, if $G(s)$ is a power series with nonnegative coefficients such that $G(1) = 1$, then G is the p.g.f. of some integer valued random variable X , which is finite with probability 1.

(4) Example Let $G(s) = (a + bs)/(1 - cs)$. When is G the p.g.f. of a finite integer valued random variable X ?

Solution First, we note that if X is finite then $G(1) = 1$, and so $a + b + c = 1$. Now we need to consider various cases.

(i) If $0 \leq c < 1$, then we can write, for any n ,

$$\begin{aligned} G(s) &= (a + bs)(1 + cs + \cdots + (cs)^n) + \frac{a + bs}{1 - cs}(cs)^{n+1} \\ &= a + (b + ac)s + (b + ac)cs^2 + \cdots + (b + ac)c^{n-1}s^n + bc^n s^{n+1} + \frac{a + bs}{1 - cs}(cs)^{n+1}. \end{aligned}$$

For $|s| < c^{-1}$, we can let $n \rightarrow \infty$ to obtain a series expansion of $G(s)$. This has the required properties of a p.g.f. if $1 \geq a \geq 0$ and $1 \geq b + ac \geq 0$. In this case, X is a nonnegative random variable.

- (ii) If $c = 1$, then $a = -b = 1$. In this case X is zero with probability 1.
 (iii) If $c > 1$, then we can use a method similar to that of (i) to obtain a different series expansion of $G(s)$, that is:

$$G = -\frac{b}{c} - \frac{ac + b}{c^2 s} - \frac{ac + b}{c^3 s^2} - \cdots$$

This is a p.g.f. if $-c \leq b \leq 0$ and $-c^2 \leq b + ac \leq 0$. In this case, X is nonpositive.

- (iv) If $c < 0$, then $a = 1$ and $b = c$. In this case, X is zero with probability 1.

See Example 14 for more insight into the nature of this probability generating function. ●

Another useful theorem is 3.6.7, which we restate here.

- (5) **Theorem** Let X be a random variable with mass function $f(k)$, and suppose that $a \leq X \leq b$. Let

$$t_n = \mathbf{P}(X > n) = \sum_{k=n+1}^b f(k).$$

Define the tail generating function

$$T(s) = \sum_{n=a}^{b-1} s^n t_n. \text{ Then, whenever both sides exist,}$$

$$(6) \quad (1-s)T(s) = s^a - G(s).$$

In particular, if $X \geq 0$, then

$$(7) \quad (1-s)T(s) = 1 - G(s).$$

Proof The left-hand side of (7) may be written as

$$\begin{aligned} (1-s) \sum_{n=a}^{b-1} \mathbf{P}(X > n) s^n &= \sum_{n=a}^{b-1} s^n \mathbf{P}(X > n) - \sum_{n=a}^{b-1} s^{n+1} \mathbf{P}(X > n) \\ &= \sum_{n=a+1}^{b-1} s^n (\mathbf{P}(X > n) - \mathbf{P}(X > n-1)) + s^a \mathbf{P}(X > a) \\ &\quad - s^b \mathbf{P}(X > b-1) \\ &= s^a - \sum_a^b \mathbf{P}(X = n) s^n = s^a - G(s), \end{aligned}$$

as required. ■

- (8) **Example 2 Revisited** Here X is uniform on $\{-a, \dots, b\}$, and so

$$\begin{aligned} T_X(s) &= \frac{s^{-a}}{1-s} - \frac{(s^{-a} - s^{b-1})}{(a+b+1)(1-s)^2} \\ &= \frac{(1-s)(a+b)s^{-a} + s^{b+1} - s^{-a+1}}{(a+b+1)(1-s)^2}. \end{aligned}$$

More generally, we can show that the identity (7) holds for unbounded nonnegative random variables. One way of doing this is to observe that the coefficients of s^n on each side are equal for all n , and then use a standard theorem about power series. ●

- (9) **Example** Let X be geometric with mass function $f(k) = (1-q)q^{k-1}; k \geq 1, 0 < q < 1$. Then

$$G(s) = \sum_{k=1}^{\infty} (1-q)q^{k-1}s^k = \frac{(1-q)s}{1-qs},$$

and

$$(1-s)T(s) = 1 - \frac{(1-q)s}{1-qs} = \frac{1-s}{1-qs}.$$

Thus,

$$T(s) = \frac{1}{1-qs}. \quad \bullet$$

For future reference, we record the following trivial corollary of (7); that is, if $\mathbf{P}(0 \leq X < \infty) = 1$, then

$$(10) \quad \sum_{j=0}^{\infty} s^j \mathbf{P}(X \leq j) = \frac{G_X(s)}{1-s}.$$

It is useful to bear in mind that conditional probability mass functions also have generating functions. Thus, if A is some event, we can write $\mathbf{P}(X = k|A) = f(k|A)$ and define the generating function

$$G_{X|A}(s) = \sum_k f(k|A)s^k = \mathbf{E}(s^X|A),$$

in the usual notation.

If $(A_i; i \geq 1)$ is a collection of disjoint events with $\bigcup_i A_i = \Omega$, then it is easy to show that

$$(11) \quad \mathbf{E}(s^X) = \sum_i \mathbf{E}(s^X|A_i)\mathbf{P}(A_i).$$

This result is often useful in finding $\mathbf{E}(s^X)$.

If the random variables X and Y are jointly distributed, then in like manner we have

$$(12) \quad \mathbf{E}(s^X|Y = y) = \sum_k s^k \mathbf{P}(X = k|Y = y) = \sum_k s^k \frac{f(k, y)}{f_Y(y)}; \quad f_Y(y) > 0.$$

As y runs over all the possible values of Y , this yields the conditional p.g.f. of X given Y

$$(13) \quad G_{X|Y}(s) = \mathbf{E}(s^X|Y).$$

We therefore have the useful result:

$$G_X(s) = \mathbf{E}(G_{X|Y}(s)) = \mathbf{E}(\mathbf{E}(s^X|Y)).$$

(14) Example 4 Revisited Suppose we have two biased coins; the first shows a head with probability a , and the second shows a head with probability $1 - c$. The first coin is tossed and, if it shows a tail then the second coin is tossed repeatedly until a head is shown. Let X be the number of times the second coin is tossed. What is $G_X(s)$?

Solution Let H be the event that the first coin shows a head. If H occurs then $X = 0$, so $\mathbf{E}(s^X|H) = 1$. If H^c occurs, then X is geometric with $f_X(k) = (1-c)c^{k-1}$; $k \geq 1$. Hence, by Example 9,

$$\mathbf{E}(s^X|H^c) = \frac{(1-c)s}{1-cs}.$$

Therefore, by (11),

$$\mathbf{E}(s^X) = a + \frac{(1-a)(1-c)s}{1-cs} = \frac{a + (1-a-c)s}{1-cs}.$$

Looking back, we see that this is the generating function considered in Example 4, case (i). It follows that we can think of $(a + bs)/(1 - cs)$ as being the p.g.f. of a random variable X , which is either zero with probability a , or with probability $1 - a$ is a geometric random variable with parameter c . Such random variables arise quite naturally in applications. ●

(15) Example A biased coin is tossed repeatedly until the first occasion when r consecutive heads have resulted. Let X be the number of tosses required. Find $\mathbf{E}(s^X)$.

Solution We suppose that the chance of a head is p , and note that if the first i tosses are $i - 1$ heads followed by a tail, then the further number of tosses required has the same mass function as X . Hence, with an obvious notation:

$$\mathbf{E}(s^X | H^{i-1}T) = s^i \mathbf{E}(s^X); \quad 1 \leq i \leq r.$$

Also, $\mathbf{E}(s^X | H^r) = s^r$. It follows that

$$\mathbf{E}(s^X) = \sum_{i=1}^r qp^{i-1}s^i \mathbf{E}(s^X) + p^r s^r$$

and so

$$\mathbf{E}(s^X) \left(1 - qs \sum_{i=0}^{r-1} (ps)^i \right) = p^r s^r.$$

Hence,

$$\mathbf{E}(s^X) = \frac{p^r s^r (1 - ps)}{1 - s + qp^r s^{r+1}}.$$

We discover different methods for proving this later. ●

6.2 Moments and the Probability Generating Function

For the remainder of this chapter, random variables are assumed to be nonnegative unless stated otherwise. In this case, whenever $|s| \leq 1$,

$$|G_X(s)| = \left| \sum_0^{\infty} f(k)s^k \right| \leq \sum_0^{\infty} f(k)|s^k| \leq \sum_0^{\infty} f(k) = 1.$$

This simple property has enormous consequences for the p.g.f. $G(s)$. These are fully explored in textbooks on calculus and analysis, so we merely state the most important relevant results here. First, we state without proof:

(1) Theorem The function $G(s)$ is differentiable for $|s| < 1$ and its derivative is

$$G'(s) = \sum_{n=1}^{\infty} nf(n)s^{n-1} < \infty.$$

At $s = 1$,

$$(2) \quad G'(1) = \lim_{s \uparrow 1} \sum_{n=1}^{\infty} n f(n) s^{n-1}$$

whether the limit is finite or not. More generally, it follows that for $k \geq 1$,

$$(3) \quad G^{(k)}(s) = \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} f(n) s^{n-k}, \quad |s| < 1$$

and

$$(4) \quad G^{(k)}(1) = \lim_{s \uparrow 1} \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} f(n) s^{n-k}.$$

Second, it follows that $G(s)$ determines the collection $\{f(k); k \geq 0\}$.

(5) **Theorem (Uniqueness)** Let X and Y have generating functions $G_X(s)$ and $G_Y(s)$. If for some $G(s)$, we have

$$G_X(s) = G_Y(s) = G(s) \text{ for } |s| < 1,$$

then X and Y have the same mass function.

Proof This follows from (3) because both $f_X(k)$ and $f_Y(k)$ are given by

$$f_X(k) = \frac{G^{(k)}(0)}{k!} = f_Y(k)$$

for all k . ■

Third, it follows that we can obtain all the moments of X from $G(s)$.

(6) **Theorem** If X has p.g.f. $G(s)$, then

$$(7) \quad \mathbf{E}(X) = G'(1);$$

more generally, the k th factorial moment is

$$(8) \quad \mu^{(k)} = \mathbf{E}(X(X-1)\dots(X-k+1)) = G^{(k)}(1);$$

and, in particular,

$$(9) \quad \text{var}(X) = G''(1) + G'(1) - (G'(1))^2.$$

Proof Equation (7) is a trivial consequence of (2), and (8) follows from (4). To see (9), write

$$(10) \quad \begin{aligned} \text{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2 = \mathbf{E}(X^2) - (G'(1))^2 \\ &= \mathbf{E}(X(X-1) + X) - (G'(1))^2, \end{aligned}$$

as required. ■

Just as (10) gives the second moment and second central moment in terms of the first two factorial moments; likewise, σ_k and μ_k may be obtained in principle in terms of $(\mu^{(k)}; k \geq 1)$.

- (11) **Example: Binomial p.g.f.** Let X have a binomial distribution with parameters n and p . Then, with $q = 1 - p$ as usual,

$$G(s) = \sum_0^n \binom{n}{k} q^{n-k} p^k s^k = (q + ps)^n.$$

Now using (8), we have

$$\mu^{(k)} = \begin{cases} \frac{n!}{(n-k)!} p^k & 1 \leq k \leq n, \\ 0 & k > n. \end{cases}$$

Hence, by (9), $\text{var}(X) = n(n-1)p^2 + np - (np)^2 = npq$. ●

- (12) **Example: Poisson p.g.f.** Let X have a Poisson distribution with parameter λ . Then

$$G(s) = \sum_0^\infty e^{-\lambda} \frac{\lambda^k}{k!} s^k = e^{+\lambda(s-1)}.$$

Hence, we find that $\mu^{(k)} = \lambda^k$, for $k \geq 1$. ●

Moments can also be obtained from the tail generating function $T(s)$ defined in Theorem 6.1.5.

- (13) **Theorem** Let X be a random variable with

$$T(s) = \sum_0^\infty s^k \mathbf{P}(X > k).$$

Then

$$(14) \quad \mathbf{E}(X) = T(1)$$

and, if $\mathbf{E}(X) < \infty$,

$$(15) \quad \text{var}(X) = 2T'(1) + T(1) - T(1)^2.$$

Proof By L'Hôpital's rule,

$$T(1) = \lim_{s \uparrow 1} \frac{1 - G(s)}{1 - s} = G'(1) = \mathbf{E}(X), \quad \text{by (7).}$$

Likewise, differentiating (6.1.7) yields

$$T'(1) = \lim_{s \uparrow 1} \left(\frac{1 - G(s)}{(1-s)^2} - \frac{G'(s)}{1-s} \right) = \frac{G''(1)}{2}, \quad \text{by L'Hôpital's rule,}$$

and the result follows using Theorem 6. ■

More generally, a straightforward extension of this theorem shows that

$$(16) \quad \mu^{(k)} = kT^{(k-1)}(1).$$

(17) **Example: Geometric p.g.f.** Let X have a geometric distribution with mass function $f(k) = (1 - q)q^{k-1}$, $k \geq 1$; $0 < q < 1$. Then, by Example 6.1.9,

$$T(s) = \frac{1}{1 - qs}.$$

Hence, by Theorem 13,

$$\mathbf{E}(X) = T(1) = \frac{1}{1 - q},$$

and likewise

$$\text{var}(X) = 2T'(1) + T(1) - T(1)^2 = \frac{q}{(1 - q)^2}.$$

From (16),

$$\mu^{(k)} = \frac{kq^{k-1}}{(1 - q)^k}.$$

We conclude this section with a note about defective probability mass functions. If X is a nonnegative random variable such that $\sum_{k=0}^{\infty} f(k) < 1$, then it still makes sense to define the generating function $G(s) = \sum_{k=0}^{\infty} s^k f(k)$. Furthermore, if $\sum_k kf(k) < \infty$, then $G'(1) = \sum_k kf(k)$. However, this is not now the expectation $\mathbf{E}(X)$, but rather the “defective” expectation

$$\mathbf{E}(XI\{X < \infty\}) = \mathbf{E}(X; X < \infty) = \mathbf{E}(X|X < \infty)\mathbf{P}(X < \infty),$$

where $I\{X < \infty\}$ is the indicator of the event that X is finite.

In the general case, we have likewise

$$G'(1) = \mathbf{E}(X||X| < \infty)\mathbf{P}(|X| < \infty)$$

when the expectation exists. In such cases, it can be of interest to calculate

$$\mathbf{E}(X|X < \infty) = \frac{G'(1)}{G(1)}.$$

6.3 Sums of Independent Random Variables

If X and Y are independent, then the mass function of their sum $Z = X + Y$ is

$$(1) \quad f_Z(k) = \sum_j f_X(j)f_Y(k - j).$$

Practical folk (such as statisticians and the like) are frequently interested in the sum of n independent random variables:

$$(2) \quad S = \sum_1^n X_i.$$

The prospect of performing the summation in (1) on $n - 1$ occasions to find $f_S(k)$ is not an attractive one. The next theorem renders it unnecessary in many important cases.

(3) **Theorem** (a) Let X_1 and X_2 be independent with generating functions $G_1(s)$ and $G_2(s)$, respectively. Then the sum $Z = X_1 + X_2$ has generating function

(4)
$$G(s) = G_1(s)G_2(s).$$

(b) More generally, if $(X_i; 1 \leq i \leq n)$ are independent with generating functions $(G_i(s); 1 \leq i \leq n)$, then the sum $Z = \sum_{i=1}^n X_i$ has generating function

$$G_Z(s) = \prod_{i=1}^n G_i(s).$$

Proof (a) Because X_1 and X_2 are independent, s^{X_1} and s^{X_2} are also independent. Hence,

$$\begin{aligned} G_Z(s) &= \mathbf{E}(s^{X_1+X_2}) = \mathbf{E}(s^{X_1})\mathbf{E}(s^{X_2}) \quad \text{by Theorem 5.3.8} \\ &= G_1(s)G_2(s). \end{aligned}$$

Part (b) is proved similarly. ■

Example: Binomial Sum Let X and Y be independent and binomially distributed with parameters (m, p) and (n, p) , respectively. Then recalling Example 6.2.11, we have

$$\begin{aligned} G_{X+Y}(s) &= \mathbf{E}(s^{X+Y}) = \mathbf{E}(s^X)\mathbf{E}(s^Y) \quad \text{by independence} \\ &= (1 - p + ps)^{m+n}. \end{aligned}$$

Hence, $X + Y$ is binomially distributed with parameters $m + n$ and p , using Theorem 6.2.5, the uniqueness theorem. ●

(5) **Example** Let $(X_i; i \geq 1)$ be independent Poisson random variables having respective parameters $(\lambda_i; i \geq 1)$. Find the mass function of $Z = \sum_{i=1}^n X_i$.

Solution Reproducing the argument of the above theorem, we have:

$$\begin{aligned} G_Z(s) &= \mathbf{E}(s^Z) = \prod_{i=1}^n \mathbf{E}(s^{X_i}) \quad \text{by independence} \\ &= \exp\left(\sum_{i=1}^n \lambda_i(s - 1)\right) \quad \text{by Example 6.2.12.} \end{aligned}$$

Thus, Z is Poisson with parameter $\sum_{i=1}^n \lambda_i$, by the uniqueness theorem. ●

- (6) **Example** Let $(X_i; i \geq 1)$ be independently and uniformly distributed on $\{1, 2, \dots, n\}$. Let $S_k = \sum_{i=1}^k X_i$, and define

$$T_n = \min \{k: S_k > n\}.$$

(Thus, T_n is the smallest number of the X_i required to achieve a sum exceeding n .) Find the mass function and p.g.f. of T_n , and hence calculate $\mathbf{E}(T_n)$ and $\text{var}(T_n)$.

Solution First, we observe that $T_n \geq j + 1$ if and only if $S_j \leq n$. Therefore,

$$(7) \quad \mathbf{P}(T_n \geq j + 1) = \mathbf{P}(S_j \leq n).$$

Now, by independence,

$$\mathbf{E}(z^{S_j}) = (\mathbf{E}(z^{X_1}))^j = \frac{1}{n^j} \left(\frac{z - z^{n+1}}{1 - z} \right)^j \quad \text{by Example 6.1.2.}$$

Hence, by Example 6.1.9,

$$(8) \quad \sum_{k=j}^{\infty} z^k \mathbf{P}(S_j \leq k) = \left(\frac{z}{n} \right)^j \frac{(1 - z^n)^j}{(1 - z)^{j+1}}.$$

Equating coefficients of z^n on each side of (8) gives

$$(9) \quad \mathbf{P}(S_j \leq n) = \frac{1}{n^j} \binom{n}{j} = \mathbf{P}(T_n \geq j + 1) \quad \text{by (7).}$$

Hence,

$$\mathbf{P}(T_n = j) = \frac{1}{n^{j-1}} \binom{n}{j-1} - \frac{1}{n^j} \binom{n}{j}.$$

From (9), T_n has tail generating function

$$\sum_{j=0}^n z^j \mathbf{P}(T_n > j) = \sum_{j=0}^n \left(\frac{z}{n} \right)^j \binom{n}{j} = \left(1 + \frac{z}{n} \right)^n.$$

Hence, from Theorem 6.2.13,

$$(10) \quad \mathbf{E}(T_n) = \left(1 + \frac{1}{n} \right)^n$$

and

$$(11) \quad \text{var}(T_n) = 2 \left(1 + \frac{1}{n} \right)^{n-1} + \left(1 + \frac{1}{n} \right)^n - \left(1 + \frac{1}{n} \right)^{2n}.$$

Finally, T_n has p.g.f.

$$(12) \quad G(z) = 1 + (z - 1) \left(1 + \frac{z}{n} \right)^n. \quad \bullet$$

Generating functions become even more useful when you are required to consider the sum of a random number of random variables.

(13) **Theorem: Random Sum** Let N and $(X_i; i \geq 1)$ be independent random variables, and suppose that N is nonnegative and that for all i

$$(14) \quad \mathbf{E}(s^{X_i}) = G(s).$$

Then the sum $Z = \sum_{i=1}^N X_i$ has generating function

$$G_Z(s) = G_N(G(s)).$$

Proof By conditional expectation,

$$\begin{aligned} \mathbf{E}(s^Z) &= \mathbf{E}(\mathbf{E}(s^Z|N)) = \mathbf{E}(\mathbf{E}(s^{X_1}) \dots \mathbf{E}(s^{X_N})) && \text{by independence} \\ &= \mathbf{E}(G(s)^N) && \text{by (14)} \\ &= G_N(G(s)) \end{aligned}$$

by Definition 6.1.3, and the result follows. ■

(15) **Example** You toss a fair coin repeatedly. Each time it shows a tail you roll a fair die, when the coin first shows a head you stop. What is the p.g.f. of the total sum of the scores shown by the rolls of the die?

Solution As you know by now, the number N of tails shown has mass function $f_N(k) = (\frac{1}{2})^{k+1}; k \geq 0$, with generating function

$$G_N(s) = \frac{1}{2-s}.$$

The score shown by each die has p.g.f.

$$G_X(s) = \frac{1}{6} \frac{s(1-s^6)}{1-s},$$

and so the p.g.f. of the total is given by Theorem 13 as

$$G(s) = \left(2 - \frac{1}{6} \frac{s(1-s^6)}{1-s} \right)^{-1}. \quad \bullet$$

(16) **Example** Let $Z = \sum_{i=1}^N X_i$, where

$$f_X(k) = \frac{-k^{-1} p^k}{\log(1-p)}; \quad k \geq 1, 0 < p < 1$$

and $f_N(k) = \lambda^k e^{-\lambda} / k!; k \geq 1, 0 < \lambda$. Show that Z has a negative binomial mass function.

Solution It is easy to show that

$$G_X(s) = \frac{\log(1-sp)}{\log(1-p)}; \quad G_N(s) = e^{\lambda(s-1)}.$$

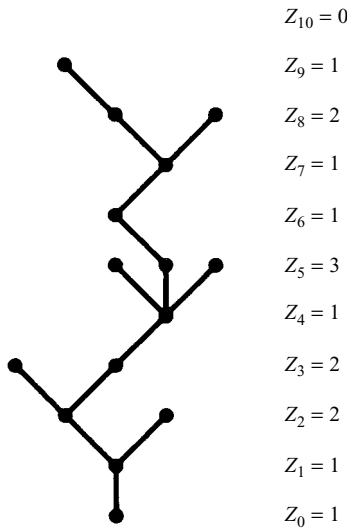


Figure 6.1 A realization of a branching process. The orientation of the diagram explains the name.

Hence,

$$G_Z(s) = e^{-\lambda} \exp(\lambda G_X(s)) = \left(\frac{1-p}{1-ps} \right)^{-\lambda(\log(1-p))^{-1}},$$

which is the p.g.f. of a negative binomial mass function. ●

(17) Example: Branching A collection of particles behaves in the following way. At time $n = 0$, there is one particle. At time $n = 1$, it is replaced by a random number X of particles, where X has mass function $f(k), k \geq 0$. At every subsequent time $n = 2, 3, \dots$, each particle in existence at that time is replaced by a random number of new particles, called its *family*. All family sizes are independent, and they all have the same mass function as the first family X . [An example is given in Figure 6.1.]

Let the number of particles in existence at time n be Z_n . Find $\mathbf{E}(s^{Z_n})$ and $\lim_{n \rightarrow \infty} \mathbf{P}(Z_n = 0)$.

Solution Let $G(s) = \mathbf{E}(s^X)$ and $G_n(s) = \mathbf{E}(s^{Z_n})$. Let the family sizes of the particles existing at time n be $(X_j; 0 \leq j \leq Z_n)$. Then we obtain the attractive and useful representation

$$Z_{n+1} = \sum_{j=0}^{Z_n} X_j$$

and, by Theorem 13,

$$G_{n+1}(s) = G_n(G(s)).$$

(This basic argument is used repeatedly in the theory of branching processes.) Hence, $G_{n+1}(s)$ is the $(n + 1)$ th iterate of $G(\cdot)$, that is to say:

$$(18) \quad G_{n+1}(s) = G(G(\dots G(s)\dots)), \quad n \geq 0.$$

Now let $\mathbf{P}(Z_n = 0) = \eta_n$, and define η to be the smallest nonnegative root of the equation

$$(19) \quad G(s) = s$$

We now show that

$$(20) \quad \lim_{n \rightarrow \infty} \eta_n = \eta.$$

First, we consider three trivial cases:

- (i) If $f(0) = 0$ then $\eta_n = G_n(0) = 0 = \eta$.
- (ii) If $f(0) = 1$ then $\eta_n = G_n(0) = 1 = \eta$.
- (iii) If $f(0) + f(1) = 1$, with $f(0)f(1) \neq 0$, then

$$\eta_n = G_n(0) = 1 - (f(1))^n \rightarrow 1 = \eta.$$

Thus, (20) is true in each case. In what follows, we exclude these cases by requiring that $0 < f(0) < f(0) + f(1) < 1$.

Now note that $\{Z_n = 0\} \subseteq \{Z_{n+1} = 0\}$ and so, by Example 1.4.11,

$$\eta_n \leq \eta_{n+1} \leq 1.$$

Hence, $\lim_{n \rightarrow \infty} \eta_n$ exists; let us denote it by λ . By (18), $G_{n+1}(0) = G(G_n(0))$; now letting $n \rightarrow \infty$ and using the continuity of $G(s)$, we find that λ is a root of (19):

$$\lambda = G(\lambda).$$

However, if for some n , $\eta_n < \eta$ then, because $G(s)$ is increasing,

$$\eta_{n+1} = G(\eta_n) \leq G(\eta) = \eta.$$

But $\eta_0 = G(0) \leq G(\eta) = \eta$, and so $\eta_n < \eta$ for all n . Hence, $\lambda \leq \eta$ and so $\lambda = \eta$. ●

Once again, we conclude with a note about defective random variables. If X and Y are defective, then they are independent if $\mathbf{P}(X = i, Y = j) = \mathbf{P}(X = i)\mathbf{P}(Y = j)$ for all finite X and Y . Hence, we can still write

$$G_{X+Y}(s) = G_X(s)G_Y(s),$$

and we can denote this by

$$\mathbf{E}(s^{X+Y}) = \mathbf{E}(s^X)\mathbf{E}(s^Y).$$

if we remember that the expectation is taken only over the finite part of the distribution.

6.4 Moment Generating Functions

The moments $(\mu_k; k \geq 1)$ of a random variable X also form a collection of real numbers, so we may also expect their generating functions to be useful. In this case, it is convenient to use the exponential generating function of the collection $(\mu_k; k \geq 1)$.

(1) **Definition** Let the random variable X have finite moments $\mu_k = \mathbf{E}(X^k)$ for all $k \geq 1$. Then the moment generating function (or m.g.f.) of X is the function $M_X(t)$ given by

$$(2) \quad M_X(t) = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}. \quad \blacktriangle$$

If X takes only a finite number of values, then we easily obtain the convenient representation

$$(3) \quad M_X(t) = \sum_{k=0}^{\infty} \mathbf{E} \frac{(X^k) t^k}{k!} = \mathbf{E} \left(\sum_{k=0}^{\infty} \frac{(Xt)^k}{k!} \right) = \mathbf{E}(e^{Xt}).$$

More generally, (3) holds provided the moments μ_k do not get too large as k increases. For example, if $\sum_{k=0}^{\infty} |\mu_k|/k! < \infty$, then $M_X(t)$ exists for $|t| < 1$, and we can use the equivalent of Theorem 6.2.1. This yields

$$(4) \quad \mu_k = M_X^{(k)}(0).$$

From (3), we find that

$$(5) \quad M_X(t) = G_X(e^t),$$

where $G_X(s)$ is the p.g.f. of X .

(6) **Example** Let X have a negative binomial distribution with mass function $f(k) = \binom{n+k-1}{k} q^k p^n, k \geq 0$. By the negative binomial expansion,

$$G(s) = \sum_{k=0}^{\infty} p^n \binom{n+k-1}{k} q^k s^k = \left(\frac{p}{1-qs} \right)^n, \quad |s| < q^{-1}.$$

Then X has moment generating function

$$M(t) = \left(\frac{p}{1-qe^t} \right)^n, \quad t < -\log q. \quad \bullet$$

Let us consider an example in which X may take negative integer values.

(7) **Example** Let X have mass function

$$f(k) = \frac{1}{2} q^{|k|-1} (1-q); \quad k = \pm 1, \pm 2, \dots,$$

where $0 < q < 1$. Then X has p.g.f.

$$\begin{aligned} G(s) &= \sum_{k \neq 0} s^k \frac{1}{2} q^{|k|-1} (1-q) \\ &= \frac{1}{2} (1-q) \left(\frac{s}{1-qs} + \frac{1}{s-q} \right), \quad q < |s| < q^{-1}. \end{aligned}$$

Hence, X has m.g.f.

$$M(t) = \frac{1}{2} (1-q) \left(\frac{e^t}{1-qe^t} + \frac{1}{e^t - q} \right), \quad \log q < t < -\log q.$$

In this case, $M(t)$ exists in an interval including the origin, and (4) holds. ●

The factorial moments ($\mu^{(k)}; k \geq 1$) also have a generating function, which is related to the probability generating function as follows:

$$\begin{aligned} (8) \quad \sum_k \frac{\mu^{(k)} t^k}{k!} &= \sum_k \mathbf{E} \frac{(X(X-1)\dots(X-k+1)) t^k}{k!} \\ &= \sum_k \sum_{n=k}^{\infty} \frac{n!}{(n-k)!} \frac{\mathbf{P}(X=n)}{k!} t^k \\ &= \sum_{n=0}^{\infty} \mathbf{P}(X=n) \sum_{k=0}^n \binom{n}{k} t^k \\ &= \sum_n \mathbf{P}(X=n) (1+t)^n = G_X(1+t). \end{aligned}$$

The change in the order of summation is justified because the terms in the sum are all nonnegative.

Now let us make the important observation that both of these moment generating functions are useful for dealing with sums of independent random variables for essentially the same reasons that made the p.g.f. so useful. To see this, let X and Y be independent, and set $Z = X + Y$. Then

$$(9) \quad M_Z(t) = \mathbf{E}(e^{t(X+Y)}) = M_X(t)M_Y(t) \quad \text{by independence.}$$

Likewise,

$$(10) \quad G_Z(1+t) = \mathbf{E}((1+t)^Z) = G_X(1+t)G_Y(1+t).$$

Finally, we record the existence of yet another function that generates the moments of X , albeit indirectly.

(11) Definition *If the function*

$$\kappa(t) = \log(\mathbf{E}(e^{Xt})) = \log(M_X(t))$$

can be expanded in powers of t , in the form

$$(12) \quad \kappa(t) = \sum_{r=1}^{\infty} \kappa_r t^r / r!,$$

then it is called the generating function of the cumulants $(\kappa_r; r \geq 1)$. ▲

(13) **Example** If X is Poisson with parameter λ , then

$$\log(M_X(t)) = \log(\exp[\lambda(e^t - 1)]) = \lambda(e^t - 1) = \sum_{r=1}^{\infty} \frac{\lambda}{r!} t^r.$$

Hence, for all r , $\kappa_r = \lambda$. ●

6.5 Joint Generating Functions

Generating functions can be equally useful when we want to consider the joint behaviour of a number of random variables. Not surprisingly, we need a joint generating function.

(1) **Definition** A random vector (X_1, \dots, X_n) , with joint mass function $f(x_1, \dots, x_n)$, has a *joint probability generating function*

$$G_X(s) = G_{X_1, \dots, X_n}(s_1, \dots, s_n) = \sum_{x_1, x_2, \dots, x_n} s^{x_1} s^{x_2} \dots s^{x_n} f(x_1, \dots, x_n). \quad \blacktriangle$$

By Theorem 5.3.1, we obtain the following useful representation of G ,

$$G_X(s) = \mathbf{E}(s^{X_1} s^{X_2} \dots s^{X_n}) = \mathbf{E}\left(\prod_{i=1}^n s_i^{X_i}\right).$$

(2) **Example** A coin shows heads with probability p or tails with probability $q = 1 - p$. If it is tossed n times, then the joint p.g.f. of the number X of heads and the number of tails is

$$\begin{aligned} G(s, t) &= \mathbf{E}(s^X t^Y) = t^n \mathbf{E}\left(\left(\frac{s}{t}\right)^X\right) = t^n \left(q + p \frac{s}{t}\right)^n \quad \text{because } X \text{ is binomial,} \\ &= (qt + ps)^n. \end{aligned} \quad \bullet$$

The fact that G is the n th power of $(qt + ps)$ suggests that independence could have been used to get this result. We use this idea in the next example.

(3) **Example: de Moivre trials** Each of a sequence of n independent trials results in a win, loss, or draw, with probabilities α , β , and γ respectively. Find the joint p.g.f. of the wins, losses, and draws, the so-called *trinomial* p.g.f.

Solution Let W_i , L_i , and D_i be the respective indicators on the i th trial of a win, loss, or draw. Then

$$\mathbf{E}(x^{W_i} y^{L_i} z^{D_i}) = \alpha x + \beta y + \gamma z.$$

But the required joint p.g.f. is

$$G(x, y, z) = \mathbf{E} \left(x^{\sum_1^n W_i} y^{\sum_1^n L_i} z^{\sum_1^n D_i} \right) = [\mathbf{E}(x^{W_i} y^{L_i} z^{D_i})]^n \quad \text{by independence,} \\ = (\alpha x + \beta y + \gamma z)^n. \quad \bullet$$

Knowledge of the joint p.g.f. entails knowledge of all the separate p.g.f.s because, for example, if X and Y have joint p.g.f. $G(s, t)$, then

$$(4) \quad G_X(s) = \mathbf{E}(s^X) = \mathbf{E}(s^X 1^Y) = G(s, 1).$$

Likewise,

$$(5) \quad G_Y(t) = G(1, t).$$

Indeed, we can quickly obtain the p.g.f. of any linear combination of X and Y ; for example, let $Z = aX + bY$, then

$$(6) \quad \mathbf{E}(s^Z) = \mathbf{E}(s^{aX+bY}) = \mathbf{E}(s^{aX} s^{bY}) = G(s^a, s^b).$$

Further, the joint generating function also provides us with the joint moments when they exist, in the same way as $G_X(s)$ provides the moments of X .

(7) **Example** Let X and Y have joint p.g.f. $G(s, t)$ and suppose that X and Y have finite variance. Then $\mathbf{E}(XY)$ exists (by the Cauchy–Schwarz inequality) and

$$\frac{\partial^2 G}{\partial s \partial t} = \frac{\partial^2}{\partial s \partial t} \mathbf{E}(s^X s^Y) = \mathbf{E}(XY s^{X-1} t^{Y-1}).$$

Hence,

$$(8) \quad \mathbf{E}(XY) = \left. \frac{\partial^2 G}{\partial s \partial t} \right|_{s=t=1}.$$

Likewise,

$$\mathbf{E}(X) = \left. \frac{\partial G}{\partial s} \right|_{s=t=1}, \quad \text{and} \quad \mathbf{E}(Y) = \left. \frac{\partial G}{\partial t} \right|_{s=t=1}.$$

Quite often, we write $G_{st}(s, t)$ for $\partial^2 G / \partial s \partial t$, and so on; in this form, the covariance of X and Y is given by

$$(9) \quad \text{cov}(X, Y) = G_{st}(1, 1) - G_s(1, 1)G_t(1, 1). \quad \bullet$$

(10) **Example 5.11 Revisited: Golf** Recall that you play n holes of golf, each of which you independently win, lose, or tie, with probabilities p , q , and r , respectively. The numbers of wins, losses, and ties are X , Y , and Z , respectively, with $X + Y + Z = n$.

(a) Find $\rho(X, Y)$.

(b) Find $\text{var}(X - Y)$.

Solution (a) By Example 3 above, we calculate

$$\mathbf{E}(x^X y^Y z^Z) = (px + qy + rz)^n = G(x, y, z) \quad \text{say.}$$

Hence,

$$\mathbf{E}(X) = G_x(1, 1, 1) = np,$$

and

$$\text{var}(X) = G_{xx}(1, 1, 1) + G_x(1) - (G_x(1))^2 = np(1 - p),$$

and

$$\mathbf{E}(XY) = G_{xy}(1, 1, 1) = n(n - 1)pq.$$

Therefore, the correlation between X and Y is

$$\begin{aligned} (11) \quad \rho(X, Y) &= \frac{\text{cov}(X, Y)}{(\text{var}(X)\text{var}(Y))^{\frac{1}{2}}} \\ &= \frac{n(n - 1)pq - n^2pq}{(n^2p(1 - p)q(1 - q))^{\frac{1}{2}}} = - \left(\frac{pq}{(1 - p)(1 - q)} \right)^{\frac{1}{2}}. \end{aligned}$$

You should compare the labour in this calculation with the more primitive techniques of Example 5.11.

(b) Using (6) with $a = 1$, $b = -1$, we have, on setting $W = X - Y$,

$$G_W(s) = \mathbf{E}(s^{X-Y}) = G(s, s^{-1}, 1) = (ps + qs^{-1} + r)^n.$$

Hence, $dG_W/ds = n(p - qs^{-2})(ps + qs^{-1} + r)^{n-1}$, and $\frac{d^2G_W}{ds^2} = n(n - 1)(p - qs^{-2})^2 (ps + qs^{-1} + r)^{n-2} + 2nqs^{-3}(ps + qs^{-1} + r)^{n-1}$.

Therefore,

$$\begin{aligned} \text{var}(W) &= n(n - 1)(p - q)^2 + 2nq + n(p - q) - n^2(p - q)^2 \\ &= n(p + q - (p - q)^2). \end{aligned} \quad \bullet$$

Finally, we record that joint generating functions provide a useful characterization of independence.

(12) **Theorem** *Let X and Y have joint p.g.f. $G(s, t)$. Then X and Y are independent if and only if*

$$(13) \quad G(s, t) = G(s, 1)G(1, t).$$

Proof If (13) holds, then equating coefficients of $s^j t^k$ gives

$$\mathbf{P}(X = j, Y = k) = \mathbf{P}(X = j)\mathbf{P}(Y = k),$$

as required. The converse is immediate by Theorem 5.3.8. ■

(14) **Example 5.5.8 Revisited: Eggs** Recall that the number X of eggs is Poisson with parameter λ , and eggs hatch independently with probability p . Let Y be the number that do hatch, and Z the number that do not. Show that Y and Z are independent, and also that $\rho(X, Y) = \sqrt{p}$.

Solution Conditional on $X = x$, the number Y of hatchings is binomial with p.g.f.

$$(15) \quad \mathbf{E}(s^Y | X = x) = (ps + 1 - p)^x.$$

Hence, by conditional expectation,

$$\begin{aligned} \mathbf{E}(y^Y z^Z) &= \mathbf{E}(y^Y z^{X-Y}) = \mathbf{E}\left(z^X \mathbf{E}\left(\left(\frac{y}{z}\right)^Y | X\right)\right) \\ &= \mathbf{E}\left(z^X \left(\frac{py}{z} + 1 - p\right)^X\right) \quad \text{by (15)} \\ &= \exp(\lambda(py + (1-p)z - 1)) \quad \text{since } X \text{ is Poisson,} \\ &= e^{\lambda p(y-1)} e^{\lambda(1-p)(z-1)}. \end{aligned}$$

Hence, Y and Z are independent by Theorem 12. Furthermore, we see immediately that Y is Poisson with parameter λp .

To find $\rho(X, Y)$, we first find the joint p.g.f. of X and Y , again using conditional expectation. Thus,

$$\mathbf{E}(s^X y^Y) = \mathbf{E}(s^X \mathbf{E}(y^Y | X)) = \mathbf{E}(s^X (py + 1 - p)^X) = \exp(\lambda(s(py + 1 - p) - 1)).$$

Hence, using (7), $\mathbf{E}(XY) = \lambda^2 p + \lambda p$, and so, using the first part, $\rho(X, Y) = \sqrt{p}$. You should compare this with the method of Example 5.5.8. ●

(16) **Example: Pairwise Independence** Independent random variables X and Y each take the values $+1$ or -1 only, and $\mathbf{P}(X = 1) = a$, with $\mathbf{P}(Y = 1) = b$. Let $Z = XY$. Show that there are values of a and b such that X , Y , and Z are pairwise independent.

(17) **Solution** Consider the joint probability generating function of X and Z .

$$\begin{aligned} G(s, t) &= \mathbf{E}(s^X t^Z) = \mathbf{E}(s^X t^{XY}) = \mathbf{E}(\mathbf{E}(s^X t^{XY} | X)) \\ &= \mathbf{E}\left(bs^X t^X + (1-b)\frac{s^X}{t^X}\right) \\ &= a\left(bst + (1-b)\frac{s}{t}\right) + (1-a)\left(\frac{b}{st} + (1-b)\frac{t}{s}\right) \\ &= \frac{abs^2 t^2 + a(1-b)s^2 + (1-a)(1-b)t^2 + b(1-a)}{st} \\ &= \frac{1}{st}\left(as^2(bt^2 + 1 - b) + \frac{(1-a)(1-b)}{b}(bt^2 + 1 - b)\right. \\ &\quad \left.+ b(1-a) - (1-b)^2(1-a)b^{-1}\right), \end{aligned}$$

which factorizes into a product of a function of s and a function of t if

$$b^2 - (1-b)^2 = 0,$$

that is if $b = \frac{1}{2}$. In this case, X and Z are independent. If $a = \frac{1}{2}$ then Y and Z are independent, and $a = b = \frac{1}{2}$ entails the pairwise independence of X , Y , and Z . ●

6.6 Sequences

In Section 4.5, we defined the convergence of a sequence of mass functions. This can be usefully connected to the convergence of corresponding sequences of generating functions. For sequences of probability generating functions, we have the following result, which we give without proof.

- (1) **Theorem** *Let $f(k)$ be a probability mass function with generating function*

$$G(s) = \sum_0^{\infty} s^k f(k),$$

and suppose that for each $n \geq 1$, $f_n(k)$ is a probability mass function with generating function

$$G_n(s) = \sum_0^{\infty} s^k f_n(k).$$

Then, as $n \rightarrow \infty$, $f_n(k) \rightarrow f(k)$ for all k , if and only if $G_n(s) \rightarrow G(s)$ for all $0 < s < 1$.

We now use this to prove a result, which we have already shown by more primitive methods.

- (2) **Example** *Let $(X_n; n \geq 1)$ be a sequence of random variables such that X_n has a binomial distribution with parameters n and λ/n , $\lambda > 0$. Then*

$$\mathbf{E}(s^{X_n}) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}s\right)^n \rightarrow e^{\lambda(s-1)} \quad \text{as } n \rightarrow \infty.$$

This is the p.g.f. of a Poisson random variable, and so as $n \rightarrow \infty$,

$$\mathbf{P}(X_n = k) \rightarrow e^{-\lambda} \lambda^k / k! \quad \bullet$$

It is often convenient to work with distributions and moment generating functions. In this case the following result (for which we give no proof) is useful.

- (3) **Theorem Continuity** *Let $\{F_n(x); n \geq 1\}$ be a sequence of distribution functions with corresponding moment generating functions $\{M_n(t) : n \geq 1\}$. If $F(x)$ is a distribution having corresponding moment generating function $M(t)$, then, as $n \rightarrow \infty$, $M_n(t) \rightarrow M(t)$ for all t , if and only if $F_n(x) \rightarrow F(x)$, whenever $F(x)$ is continuous. \bullet*

Additional conditions are required to link the convergence of a sequence of mass functions or distributions and the convergence of their moments. The following theorem (for which again we offer no proof) is for a sequence of distributions.

- (4) **Theorem** *Suppose that for each $n \geq 1$, the distribution $F_n(x)$ has moments $\{\mu_j(n); j \geq 1\}$, such that $|\mu_j(n)| < a_j < \infty$.*

(i) Let $F_n(x) \rightarrow F(x)$, as $n \rightarrow \infty$, wherever $F(x)$ is continuous. Then as $n \rightarrow \infty$, for each j ,

$$\mu_j(n) \rightarrow \mu_j < \infty,$$

and $(\mu_j; j \geq 1)$ are the moments of $F(x)$.

(ii) Conversely, for each $j \geq 1$, as $n \rightarrow \infty$, suppose that

$$\mu_j(n) \rightarrow \mu_j < \infty,$$

where $\{\mu_j; 1 \leq j\}$ are the moments of a unique distribution $F(x)$. Then, as $n \rightarrow \infty$, $F_n(x) \rightarrow F(x)$ wherever $F(x)$ is continuous.

There is a corresponding result for sequences of mass functions.

These theorems find applications (for example) in the theory of random graphs, and other combinatorial problems where moments are more tractable than distributions.

(5) **Example** Let X_n have the binomial distribution with parameters n and λ/n . Then, by Example 6.2.11,

$$\mu^{(k)} = \begin{cases} \frac{n!}{(n-k)!} \left(\frac{\lambda}{n}\right)^k & 1 \leq k \leq n \\ 0 & k > n \end{cases} \rightarrow \lambda^k,$$

as $n \rightarrow \infty$. But, by Example 6.2.12, these are the factorial moments of the Poisson distribution. Hence, as $n \rightarrow \infty$, $\mathbf{P}(X_n = k) \rightarrow e^{-\lambda} \lambda^k / k!$, which we proved directly in Example 6.6.2, and earlier in Example 4.10. ●

(6) **Example: Matching Again** Recall that we are assigning n distinct letters randomly to n matching envelopes, and X is the number of matched pairs (of letter and envelope) that result. Consider the k th factorial moment of X .

$$\mu^{(k)} = \mathbf{E}(X(X-1)\dots(X-k+1)); \quad 1 \leq k \leq n.$$

Let I_j be the indicator of the event that the j th envelope contains the matching letter. Then the sum

$$(7) \quad S = \sum_{j_1 < j_2 < \dots < j_k} I_{j_1} \dots I_{j_k}$$

is just the number of ways of choosing a set of size k from the set of matching pairs. Another way of writing this is as $\binom{X}{k}$. Hence,

$$(8) \quad \begin{aligned} \frac{\mu^{(k)}}{k!} &= \mathbf{E} \frac{(X(X-1)\dots(X-k+1))}{k!} \\ &= \mathbf{E} \left(\binom{X}{k} \right) \\ &= \mathbf{E}(S) \\ &= \binom{n}{k} \mathbf{E}(I_{j_1} \dots I_{j_k}) \end{aligned}$$

$$\begin{aligned}
&= \binom{n}{k} \mathbf{P}(\text{a given set of } k \text{ all match}) \\
&= \binom{n}{k} \left(\frac{1}{n} \cdot \frac{1}{n-1} \cdots \frac{1}{n-k+1} \right) = \frac{1}{k!}.
\end{aligned}$$

Hence,

$$\mu^{(k)} = \begin{cases} 1; & k \leq n \\ 0; & k > n, \end{cases} \rightarrow 1 \quad \text{for all } k$$

as $n \rightarrow \infty$. But these are the factorial moments of the Poisson distribution with parameter 1, and so as $n \rightarrow \infty$

$$(9) \quad \mathbf{P}(X = k) \rightarrow \frac{1}{ek!}. \quad \bullet$$

We conclude with an example that leads into the material of Chapter 7.

(10) **Example 6.3.6 Revisited** Recall that the p.g.f. of T_n (where T_n is the number of uniform random variables required to give a sum greater than n) is

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

What happens as $n \rightarrow \infty$? How do you interpret this?

Solution From (6.3.12), as $n \rightarrow \infty$,

$$\mathbf{E}(s^{T_n}) \rightarrow 1 + (s-1)e^s$$

or, equivalently,

$$\mathbf{E}(e^{tT_n}) \rightarrow 1 + (e^t - 1)e^{e^t}.$$

It follows that

$$\mathbf{P}(T_n = k) \rightarrow \frac{1}{(k-1)!} - \frac{1}{k!}.$$

The limiting factorial moments have a simple form for

$$\mathbf{E}((1+t)^{T_n}) \rightarrow 1 + te^{1+t} = 1 + \sum_{k=1}^{\infty} \frac{e}{(k-1)!} t^k.$$

Hence, in the limit $\mu^{(k)} = ek, k \geq 1$.

To interpret this, we return to the original definition

$$T_n = \min \left\{ k : \sum_1^n X_i > n \right\} = \min \left\{ k : \sum_1^n \left(\frac{X_i}{n} \right) > 1 \right\},$$

where each X_i/n is uniformly distributed on $\{1/n, 2/n, \dots, 1\}$. However, the limit of this sequence of uniform mass functions as $n \rightarrow \infty$ is *not* the mass function of a discrete random variable.

Intuitively, you may feel that it is approaching the distribution of a variable that is uniformly distributed over the *interval* $[0, 1]$. This vague remark can in fact be given a meaning if we introduce new objects – namely, continuous random variables. This is the subject of the next two chapters; after much technical development, it can be shown that the limit of T_n above is indeed the number of independent random variables, each uniform on $[0, 1]$, required to produce a sum greater than 1.

6.7 Regeneration

Many interesting and important sequences of random variables arise as some process evolves in time. Often, a complete analysis of the process may be too difficult, and we seek simplifying ideas. One such concept, which recurs throughout probability, is the idea of regeneration. Here is an illustration:

- (1) **Example: Maze** You are trying to traverse an unknown labyrinth. You set off at a constant speed from the clock by the portal, and each time a decision is required you choose at random from the alternatives. It is dark, and you have no pencil and paper, so a description of the process (i.e., your route) is impossible.

However, each time you arrive back at the portal, you can look at the clock and record T_r , the time at which you return to the clock for the r th time; $T_0 = 0$ say. Now it is clear from the setup that, when you set off for a second time (at time T_1), your chance of following any given route around the maze is the same as when you set off for the first time. Thus, the time until your second return, which is $T_2 - T_1$, has the same distribution as T_1 and is independent of T_1 . The same is true of every subsequent interval between successive visits to the portal. These times ($T_n; n \geq 0$) are regenerative in the sense that the distribution of your paths starting from T_n is the same as the distribution starting from T_m , for all m and n .

Of course, if you leave pebbles at junctions, or take a ball of string, or make a map, then this is no longer true. ●

Here is another archetypal illustration.

- (2) **Example: Renewal** A machine started at $T_0 = 0$ uses a bit that wears out. As soon as it wears out, at time T_1 , it is replaced by a similar bit, which in turn is replaced at T_2 . Assuming that the machine performs much the same tasks as time passes, it seems reasonable to assume that the collection ($X_n; n \geq 1$), where $X_n = T_n - T_{n-1}$, are independent and identically distributed. The replacement (or renewal) times are regenerative. ●

In fact, we have already used this idea of restarting from scratch; see, for example, Examples 4.4.9, 4.19, and 5.4.13.

Here is another elementary illustration.

Example Three players A, B , and C take turns rolling a fair die in the order $ABCAB \dots$ until one of them rolls a 5 or a 6. Let X_0 be the duration of the game (i.e., the number of rolls). Let A_0 be the event that A wins, and let A_r be the event that A wins after the r th roll. Let $W_i, i \geq 1$ be the event that the game is won on the i th roll.

Of course,

$$A_0 = W_1 \cup \{W_1^c \cap W_2^c \cap W_3^c \cap A_3\}, \text{ where } \mathbf{P}(W_1) = \frac{1}{3}.$$

Now, as usual, we denote the indicator of any event E by $I(E)$, and so

$$\mathbf{E}(s^{X_0} I(W_1)) = \frac{1}{3}s.$$

Next we observe that if the first three rolls fail to yield 5 or 6, then the process regenerates (in the sense discussed above), so $X_0 = 3 + X_3$, where X_3 has the same distribution as X_0 . Hence,

$$\mathbf{E}(s^{X_0} I(A_3)) = \mathbf{E}(s^{3+X_3} I(A_3)) = s^3 \mathbf{E}(s^{X_0} I(A_0)).$$

Therefore, we can write

$$\begin{aligned} \mathbf{E}(s^{X_0} I(A_0)) &= \mathbf{E}(s^{X_0} \{I(W_1) + I(W_1^c \cap W_2^c \cap W_3^c \cap A_3)\}) \\ &= \frac{1}{3}s + \frac{8}{27}s^3 \mathbf{E}(s^{X_0} I(A_0)). \end{aligned}$$

Hence,

$$\mathbf{E}(s^{X_0} I(A_0)) = \frac{\frac{1}{3}s}{1 - \frac{8}{27}s^3}.$$

Likewise, in an obvious notation,

$$\mathbf{E}(s^{X_0} I(B_0)) = \frac{\frac{2}{9}s^2}{1 - \frac{8}{27}s^3},$$

and

$$\mathbf{E}(s^{X_0} I(C_0)) = \frac{\frac{4}{27}s^3}{1 - \frac{8}{27}s^3}.$$

Hence,

$$\mathbf{E}(s^{X_0}) = \mathbf{E}(s^{X_0}(I(A_0) + I(B_0) + I(C_0))) = \frac{4s^3 + 6s^2 + 9s}{27 - 8s^3}. \quad \bullet$$

Now we consider a more general case.

Let H be some phenomenon (or happening) that may occur or not at any time $n = 1, 2, 3, \dots$. Let H_n be the event that H occurs at time n , and define X_n to be the time interval between the $(n-1)$ th and n th occurrences of H . Thus,

$$X_1 = \min \{n > 0: H_n \text{ occurs}\}$$

$$X_1 + X_2 = \min \{n > X_1: H_n \text{ occurs}\}$$

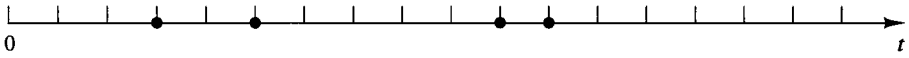


Figure 6.2 A delayed renewal process. Here, $X_1 = T_1 = 3$; $X_2 = 2$, $X_3 = 5$, $X_4 = 1, \dots$

and so on. We suppose that $(X_n; n \geq 2)$ are independent and identically distributed random variables with mass function $(f(k); k \geq 1)$ and p.g.f. $G(s)$. The first interval X_1 is independent of $(X_n; n \geq 2)$, but its mass function may or may not be the same as that of X_2 . This gives rise to two cases:

Case (O) The *ordinary* case. The mass function of X_1 is $(f(k); k \geq 1)$, the same as X_2 .

Case (D) The *delayed* case. The mass functions of X_1 is $(d(k); k \geq 1)$, and $\mathbf{E}(s^{X_1}) = D(s)$.

These two cases admit a conventional interpretation: in the ordinary case, we suppose that H_0 occurred, so X_1 has the same mass function as the other intervals; in the delayed case, H_0 did not occur, so X_1 may have a different mass function.

The mathematical structure described above is known as a *recurrent event process*, or alternatively as a *discrete renewal process*. The important point about such a process is that each time H occurs, the process regenerates itself, in the sense discussed above. Figure 6.2 displays a realization of a renewal process.

Now Examples 1 and 2 make it clear that there are two essentially different types of renewal process. In Example 1, there is always a chance that you do traverse the maze (or encounter the Minotaur), and so do not return to the entrance. That is,

$$\mathbf{P}(X_2 < \infty) < 1.$$

Such a process is called *transient*. However, all machine bits wear out eventually, so in (2) we have

$$\mathbf{P}(X_2 < \infty) = 1.$$

Such a process is called *persistent* (or recurrent).

Now define the probabilities $u_n = \mathbf{P}(H_n)$, $n \geq 1$. A natural question is to ask whether this distinction between persistence and transience can also be observed in the properties of u_n . (The answer is yes, as we will see.)

It is customary to make a further distinction between two different types of persistent renewal process.

Definition If $\mathbf{E}(X_2) = \infty$, then the process is said to be null; if $\mathbf{E}(X_2) < \infty$, then the process is said to be nonnull.

Note that $\mathbf{E}(X_2)$ is sometimes known as the mean recurrence time of the process. ▲

Ordinary renewal is just a special case of delayed renewal, of course, but it is convenient to keep them separate. We therefore define $u_0 = 1$ and $u_n = \mathbf{P}(H_n)$, in the ordinary case; and $v_0 = 0$, and $v_n = \mathbf{P}(H_n)$, in the delayed case. These have respective generating

functions, $U(s) = \sum_0^\infty u_n s^n$ and $V(s) = \sum_1^\infty v_n s^n$. [Remember that $U(s)$ and $V(s)$ are not *probability* generating functions in the sense in which we use that term.]

Now we have the following:

$$(3) \quad \text{Theorem} \quad (i) \quad U(s) = \frac{1}{1 - G(s)}$$

$$(ii) \quad V(s) = \frac{D(s)}{1 - G(s)}.$$

Proof By conditional probability, in Case (O),

$$u_n = \sum_{k=1}^n \mathbf{P}(H_n | X_1 = k) \mathbf{P}(X_1 = k).$$

However, given H_k , the probability of any later occurrence of H is as if the process started at k . That is to say

$$(4) \quad \mathbf{P}(H_n | X_1 = k) = \mathbf{P}(H_n | H_k) = \mathbf{P}(H_{n-k}) = u_{n-k}, \quad n \geq k.$$

Hence,

$$u_n = \sum_{k=1}^n u_{n-k} f(k).$$

Because the right-hand sum is a convolution, its generating function is the product of the two generating functions $U(s)$ and $G(s)$, whence

$$(5) \quad U(s) - 1 = U(s)G(s).$$

Likewise, in Case (D),

$$\begin{aligned} v_n &= \sum_{k=1}^n \mathbf{P}(H_n | X_1 = k) \mathbf{P}(X_1 = k) = \sum_{k=1}^n \mathbf{P}(H_n | X_1 = k) d(k) \\ &= \sum_{k=1}^n u_{n-k} d(k) \quad \text{by (4)}. \end{aligned}$$

Hence,

$$(6) \quad V(s) = U(s)D(s) = \frac{D(s)}{1 - G(s)} \quad \text{by (5)}. \quad \blacksquare$$

Thus, given $G(s)$ [and $D(s)$ in the delayed case] we can in principle find $\mathbf{P}(H_n)$, the probability that H occurs at time n , by expanding $U(s)$ in powers of s .

Conversely, given $V(s)$ [and $U(s)$ in the delayed case] we can find $G(s)$ and $D(s)$, and also decide whether the process is transient or persistent.

(7) **Corollary** If $U(1) < \infty$, then the process is transient. Otherwise, it is persistent.

Proof This follows immediately from Theorem 3(i). \blacksquare

- (8) **Example: Coincidences** Suppose that a number c of independent simple symmetric random walks are started simultaneously from the origin. Let H be the “happening” that they are all at the origin, so H_{2n} is the event that all the c walks are at 0 on the $2n$ th step. Show that H is persistent when $c = 2$, but H is transient for $c \geq 3$.

Solution For $c = 2$, we recall that

$$\begin{aligned} \mathbf{P}(H_{2n}) &= (u_{2n})^2 = \left(\frac{1}{4^n} \binom{2n}{n} \right)^2 \\ &= \left(1 - \frac{1}{2} \right)^2 \left(1 - \frac{1}{4} \right)^2 \cdots \left(1 - \frac{1}{2n} \right)^2 \\ &> \left(1 - \frac{1}{2} \right)^2 \left(1 - \frac{1}{3} \right) \left(1 - \frac{1}{4} \right) \cdots \left(1 - \frac{1}{2n-1} \right) \left(1 - \frac{1}{2n} \right) \\ &= \frac{1}{4n} \quad \text{on successive cancellation.} \end{aligned}$$

Hence, H is persistent as $\sum_n \mathbf{P}(H_{2n})$ diverges.

For $c \geq 3$, we have similarly that

$$\begin{aligned} \mathbf{P}(H_{2n}) &= \left(\frac{1}{4^n} \binom{2n}{n} \right)^c = \left(\left(1 - \frac{1}{2} \right)^2 \left(1 - \frac{1}{4} \right)^2 \cdots \left(1 - \frac{1}{2n} \right)^2 \right)^{c/2} \\ &< \left(\frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdot \frac{4}{5} \cdots \frac{2n-1}{2n} \cdot \frac{2n}{2n+1} \right)^{c/2} = \left(\frac{1}{2n+1} \right)^{c/2}. \end{aligned}$$

Hence, H is transient as $\sum_n (1/(2n+1))^{c/2} < \infty$. ●

- (9) **Example: Stationary Renewal** Let $X > 0$ have p.g.f. $G(s)$. Show that if $\mathbf{E}(X) < \infty$, then

$$H(s) = \frac{1}{\mathbf{E}(X)} \frac{1 - G(s)}{1 - s}$$

is the p.g.f. of a nonnegative random variable.

Now consider the delayed case of a recurrent event process in which $\mathbf{E}(s^{X_2}) = G(s)$ and $\mathbf{E}(s^{X_1}) = H(s)$. Show that for all n

$$(10) \quad \mathbf{P}(H_n) = \frac{1}{G'(1)}.$$

Solution From (6.1.6), we have that $H(s)$ is a power series with nonnegative coefficients. Furthermore, by L'Hôpital's rule,

$$H(1) = \lim_{s \uparrow 1} \frac{-G'(s)}{-\mathbf{E}(X)} = 1$$

Hence, $H(s)$ is a p.g.f. Finally, if $D(s) = H(s)$ in (6), then

$$V(s) = \frac{1}{\mathbf{E}(X_2)(1-s)},$$

and the result follows. ●

6.8 Random Walks

Recall that if $(X_i; i \geq 1)$ are independent and identically distributed, then $S_n = S_0 + \sum_1^n X_i$ is a random walk. Because generating functions have been so useful in handling sums of random variables, we may expect them to be exceptionally useful in analysing random walks. If X has p.g.f. $G(z)$, then trivially we have, when $S_0 = 0$.

$$G_n(z) = \mathbf{E}(z^{S_n}) = (G(z))^n.$$

It follows that we can define the function H by

$$(1) \quad H(z, w) = \sum_{n=0}^{\infty} w^n G_n(z) = (1 - wG(z))^{-1}.$$

This bivariate generating function tells us everything about S_n in principle, as $\mathbf{P}(S_n = r)$ is the coefficient of $z^r w^n$ in $H(z, w)$. However, the analytical effort required to work at this level of generality is beyond our scope. We proceed by considering simple examples.

- (2) **Example: Simple Symmetric Random Walk** Let $(S_n = \sum_1^n X_i; n \geq 0)$ be a simple symmetric random walk, with $S_0 = 0$. Let H_n be the event that $S_n = 0$. Because steps of the walk are independent and identically distributed, it follows that visits to the origin form an ordinary renewal process. Here, $u_n = \mathbf{P}(S_n = 0)$.

Define the first passage times,

$$T_j = \min \{n > 0 : S_n = j | S_0 = 0\},$$

and the generating functions, $U(s) = \sum_0^{\infty} u_n s^n$ and $G_j(s) = \mathbf{E}(s^{T_j})$.

Find $U(s)$ and $G_0(s)$, and show that the simple symmetric random walk is persistent null.

Solution We give two methods of finding $U(s)$ and $G_j(s)$. For the first, define

$$T_j^* = \min \{n : S_n = j - 1 | S_0 = -1\}$$

and let \hat{T}_1 be a random variable having the same distribution as T_1 , but independent of T_1 . Because the steps of the walk are independent, and symmetrically and identically distributed, it follows that

$$(3) \quad \mathbf{E}(s^{T_1}) = \mathbf{E}(s^{T_1^*}),$$

$$(4) \quad T_2 = T_1 + \hat{T}_1,$$

and

$$(5) \quad \mathbf{E}(s^{T_2^*}) = \mathbf{E}(s^{T_2}) = \mathbf{E}(s^{T_1 + \hat{T}_1})$$

$$(6) \quad = (G_1(s))^2, \quad \text{by independence.}$$

Hence, by conditional expectation,

$$(7) \quad \begin{aligned} G_1(s) &= \mathbf{E}(\mathbf{E}(s^{T_1}|X_1)) = \frac{1}{2}\mathbf{E}(s^{T_1}|X_1 = 1) + \frac{1}{2}\mathbf{E}(s^{T_1}|X_1 = -1) \\ &= \frac{1}{2}s + \frac{1}{2}\mathbf{E}(s^{1+T_2^*}) = \frac{1}{2}s + \frac{1}{2}s(G_1(s))^2 \quad \text{by (6)}. \end{aligned}$$

One root of (7) is a probability generating function, so this root is $G_1(s)$, namely,

$$G_1(s) = (1 - (1 - s^2)^{\frac{1}{2}})/s.$$

Now, using conditional expectation again,

$$(8) \quad \begin{aligned} G_0(s) &= \mathbf{E}(\mathbf{E}(s^{T_0}|X_1)) \\ &= \frac{1}{2}\mathbf{E}(s^{T_0}|X_1 = 1) + \frac{1}{2}\mathbf{E}(s^{T_0}|X_1 = -1) \\ &= \frac{1}{2}s\mathbf{E}(s^{T_1}) + \frac{1}{2}s\mathbf{E}(s^{T_1}) = sG_1(s) \quad \text{by (3)} \\ &= 1 - (1 - s^2)^{\frac{1}{2}}. \end{aligned}$$

Hence,

$$(9) \quad U(s) = (1 - s^2)^{-1/2} \quad \text{by Theorem 6.7.3.}$$

Alternatively, we could observe that $S_{2n} = 0$ if and only if the walk has taken n steps in each direction. They may be taken in any order so

$$(10) \quad u_{2n} = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}.$$

Now recall that by the negative binomial theorem

$$(11) \quad \sum_0^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} x^n = (1 - x)^{-\frac{1}{2}},$$

and (9) and (8) follow.

Setting $s = 1$ shows that $G_0(1) = 1$ (and $U(1) = \infty$) so H is persistent. However,

$$\frac{d}{ds}G_0(s) = \frac{s}{(1 - s^2)^{\frac{1}{2}}},$$

and setting $s = 1$ shows that H is null; the expected number of steps to return to the origin is infinite as we know already, recall Example 5.6.27. ●

Now that we have the generating functions $U(s)$ and $G_0(s)$, we can provide slicker derivations of earlier results. For example,

$$G_0(s) = 1 - (1 - s^2)^{\frac{1}{2}} = 1 - (1 - s^2)U(s).$$

Hence, equating coefficients of s^{2k} gives (5.18.1)

$$(12) \quad f_{2k} = u_{2k-2} - u_{2k}.$$

Also,

$$s \frac{d}{ds} G_0(s) = s^2(1-s^2)^{-\frac{1}{2}} = s^2 U(s)$$

and so equating coefficients again gives (5.18.6)

$$(13) \quad 2k f_{2k} = u_{2k-2}.$$

See Problem 41 for another simple application of this. Here is a trickier application.

(14) Example: Truncated Walk Let $(S_n; n \geq 1)$ be a simple symmetric random walk with $S_0 = 0$, and let $T = \min\{n > 0 : S_n = 0\}$. Let $T \wedge 2m = \min\{T, 2m\}$ and show that

$$(15) \quad \mathbf{E}(T \wedge 2m) = 4mu_{2m} = 2\mathbf{E}(|S_{2m}|).$$

Solution We establish (15) by showing that all three terms have the same generating function. Equality then follows by the uniqueness theorem. First,

$$\sum_0^\infty 4mu_{2m} s^{2m} = 2s \sum_1^\infty 2ms^{2m-1} u_{2m} = 2s \frac{d}{ds} U(s) = \frac{2s^2}{(1-s^2)^{3/2}}.$$

Second, recalling (13) and (5.18.2),

$$\mathbf{E}(T \wedge 2m) = \sum_{k=1}^m 2k f_{2k} + 2m \mathbf{P}(T > 2m) = \sum_{k=1}^m u_{2k-2} + 2mu_{2m}.$$

Hence,

$$\begin{aligned} \sum s^{2m} \mathbf{E}(T \wedge 2m) &= \sum_m s^{2m} \sum_{k=0}^{m-1} u_{2k} + s \sum_m 2ms^{2m-1} u_{2m} \\ &= \frac{s^2 U(s)}{1-s^2} + sU'(s) \quad \text{using (6.1.10)} \\ &= \frac{2s^2}{(1-s^2)^{\frac{3}{2}}}. \end{aligned}$$

Finally, using the hitting time theorem (5.6.17),

$$\begin{aligned} \sum_m s^{2m} \mathbf{E}(|S_{2m}|) &= 2 \sum_m s^{2m} \sum_{k=1}^m 2k \mathbf{P}(S_{2m} = 2k) \\ &= 2 \sum_m s^{2m} \sum_{k=1}^m 2m f_{2k}(2m) \\ &= 2s \frac{d}{ds} \sum_m s^{2m} \sum_{k=1}^m f_{2k}(2m) \end{aligned}$$

$$\begin{aligned}
&= 2s \frac{d}{ds} (G_2(s) + G_4(s) + G_6(s) + \dots) \\
&= 2s \frac{d}{ds} \frac{(G_1(s))^2}{(1 - (G_1(s))^2)} \quad \text{by Example (2)} \\
&= s \frac{d}{ds} ((1 - s^2)^{-\frac{1}{2}} - 1) = \frac{s^2}{(1 - s^2)^{\frac{3}{2}}}.
\end{aligned}$$

As a final example of the use of generating functions in random walks, we establish yet another arc-sine law.

(16) Example: Arc-Sine Law for Leads Let $(S_n; n \geq 0)$ be a simple symmetric random walk with $S_0 = 0$. Of the first $2n$ steps, let L_{2n} denote the number that do not enter the negative half-line. Show that

$$(17) \quad \mathbf{P}(L_{2n} = 2k) = 4^{-n} \binom{2k}{k} \binom{2n - 2k}{n - k}.$$

Solution Define the generating functions

$$(18) \quad G_{2n}(s) = \mathbf{E}(s^{L_{2n}}) \quad \text{and} \quad H(s, t) = \sum_{n=0}^{\infty} t^{2n} G_{2n}(s).$$

Let T be the number of steps until the walk first revisits zero, and recall that

$$(19) \quad F(s) = \mathbf{E}(s^T) = \sum_1^{\infty} s^{2r} f(2r) = 1 - (1 - s^2)^{\frac{1}{2}}.$$

Now using conditional expectation

$$\begin{aligned}
\mathbf{E}(s^{L_{2n}}) &= \mathbf{E}(\mathbf{E}(s^{L_{2n}} | T)) \\
&= \sum_{r=1}^n \mathbf{E}(s^{L_{2n}} | T = 2r) f(2r) + \mathbf{E}(s^{L_{2n}} | T > 2n) \sum_{r=n+1}^{\infty} f(2r).
\end{aligned}$$

Now, depending on the first step,

$$\mathbf{P}(L_T = T) = \frac{1}{2} = \mathbf{P}(L_T = 0),$$

and visits to zero constitute regeneration points for the process L_{2n} . Hence, we may rewrite (18) as

$$G_{2n}(s) = \sum_{r=1}^n G_{2n-2r} \left(\frac{1}{2}(s^{2r} + 1) \right) f(2r) + \sum_{r=n+1}^{\infty} f(2r) \left(\frac{1}{2}(s^{2n} + 1) \right).$$

Multiplying by t^{2n} and summing over n yields

$$H(s, t) = \frac{1}{2} H(s, t)(F(st) + F(t)) + \frac{1}{2} \cdot \frac{1 - F(st)}{1 - t^2 s^2} + \frac{1}{2} \cdot \frac{1 - F(t)}{1 - t^2},$$

by the convolution theorem. Now substituting for $F(\cdot)$ from (19) gives

$$H(s, t) = ((1 - s^2 t^2)(1 - t^2))^{-\frac{1}{2}}.$$

The coefficient of $t^{2n}s^{2k}$ in this is (17). Now use Exercise 5.18.8 to produce the arc-sine distribution. ●

6.9 Review and Checklist for Chapter 6

All random variables have a probability distribution, and many of them also have moments. In this chapter, we introduced two miraculous devices to help us with many of the chores involved in handling and using probabilities and moments.

Probability generating function of X : p.g.f.

$$G_X(s) = \sum_n \mathbf{P}(X = n)s^n = \mathbf{E}s^X.$$

Moment generating function of X : m.g.f.

$$M_X(t) = \sum_n \mathbf{P}(X = n)e^{nt} = \mathbf{E}e^{tX}.$$

You can think of these functions as organizers that store a collection of objects that they will regurgitate on demand. Remarkably, they will often produce other information if suitably stimulated; thus, the p.g.f. will produce the moments (if any exist), and the m.g.f. will produce the probability distribution (in most cases). We used them to study sums of independent random variables, branching processes, renewal theory, random walks, and limits. They have these properties:

Connections:

$$M_X(t) = G_X(e^t) \quad \text{and} \quad G_X(s) = M_X(\log s).$$

Tails:

$$T(s) = \sum_{n=0}^{\infty} s^n \mathbf{P}(X > n) = \frac{1 - G_X(s)}{1 - s}, \quad \text{when } X \geq 0.$$

Uniqueness:

$$f_X(k) = G_X^{(k)}(0)/k!.$$

Moments:

$$\begin{aligned} \mu^{(k)} &= G_X^{(k)}(1) \quad \text{and} \quad \mathbf{E}X^k = M_X^{(k)}(0) \\ \mu^{(k)} &= kT^{(k-1)}(1). \end{aligned}$$

[Where $\mu^{(k)}$ is the k th factorial moment.]

Sums and random sums: For independent $(X_n; n \geq 1)$,

$$\mathbf{E}s^{X_1 + \cdots + X_n} = \prod_{r=1}^n G_{X_r}(s); \quad \mathbf{E} \exp(t(X_1 + \cdots + X_n)) = \prod_{r=1}^n M_{X_r}(t).$$

For independent $(X_n; n \geq 1)$ and independent nonnegative integer valued N ,

$$\mathbf{E}s^{X_1 + \cdots + X_N} = G_N(G_X(s)); \quad \mathbf{E} \exp(t(X_1 + \cdots + X_N)) = G_N(M_X(t)).$$

Joint generating functions:

$$G_{\mathbf{X}}(\mathbf{s}) = \mathbf{E} \prod_{r=1}^n s_r^{X_r}$$

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbf{E} \left[\exp \left(\sum_{r=1}^n t_r X_r \right) \right]$$

Independence: X and Y are independent if and only if

$$\mathbf{E}(x^X y^Y) = G_X(x)G_Y(y), \quad \text{for all } x, y, \text{ or}$$

$$\mathbf{E}[\exp(sX + tY)] = M_X(s)M_Y(t), \quad \text{for all } s, t.$$

Branching: If the family size p.g.f. is $G(s)$, then

$$G_{m+n}(s) = \mathbf{E}s^{Z_{m+n}} = G_m(G_n(s)) = G_n(G_m(s)).$$

The probability η of ultimate extinction is the smallest positive root of $G(x) = x$.

Special generating functions:

Binomial distribution $(q + ps)^n$

Uniform distribution on $\{0, 1, \dots, n\}$ $\frac{1-s^{n+1}}{(n+1)(1-s)}$

Geometric distribution $\frac{ps}{1-qs}$

Poisson distribution $e^{\lambda(s-1)}$

Negative binomial distribution $(\frac{ps}{1-qs})^n$

Logarithmic distribution $\frac{\log(1-sp)}{\log(1-p)}$

Checklist of Terms for Chapter 6

- 6.1 probability generating function
 - tail generating function
- 6.2 uniqueness theorem
 - factorial moments
- 6.3 sums and random sums
 - branching process
 - extinction probability
- 6.4 moment generating function
 - cumulant generating function
- 6.5 joint probability generating function
 - factorization and independence
- 6.6 continuity theorem
- 6.7 renewal process
 - persistent
 - transient
 - null, nonnull
- 6.8 simple random walk
 - arc-sine law for leads

Finally, we note that we have occasionally used elementary ideas from calculus in this chapter, and we need to do so more frequently in Chapter 7. We therefore include a brief synopsis of the basic notions.

Appendix: Calculus

Fundamental to calculus is the idea of taking limits of functions. This in turn rests on the idea of convergence.

Convergence Let $(x_n; n \geq 1)$ be a sequence of real numbers. Suppose that there is a real number a such that $|x_n - a|$ is always ultimately as small as we please; formally,

$$|x_n - a| < \epsilon \quad \text{for all } n > n_0,$$

where ϵ is arbitrarily small and n_0 is finite. ▲

In this case, the sequence (x_n) is said to converge to the limit a . We write *either*

$$x_n \rightarrow a \quad \text{as } n \rightarrow \infty \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = a.$$

Now let $f(x)$ be any function defined in some interval (α, β) , except possibly at the point $x = a$.

Let (x_n) be a sequence converging to a , such that $x_n \neq a$ for any n . Then $(f(x_n); n \geq 1)$ is also a sequence; it may converge to a limit l .

Limits of Functions If the sequence $(f(x_n))$ converges to the same limit l for every sequence (x_n) converging to a , $x_n \neq a$, then we say that the limit of $f(x)$ at a is l . We write *either*

$$f(x) \rightarrow l \quad \text{as } x \rightarrow a, \quad \text{or} \quad \lim_{x \rightarrow a} f(x) = l. \quad \blacktriangle$$

Suppose now that $f(x)$ is defined in the interval (α, β) , and let $\lim_{x \rightarrow a} f(x)$ be the limit of $f(x)$ at a . This may or may not be equal to $f(a)$. Accordingly, we define:

Continuity The function $f(x)$ is continuous in (α, β) if, for all $a \in (\alpha, \beta)$,

$$\lim_{x \rightarrow a} f(x) = f(a). \quad \blacktriangle$$

Now, given a continuous function $f(x)$, we are often interested in two principal questions about $f(x)$.

- (i) What is the slope (or gradient) of $f(x)$ at the point $x = a$?
- (ii) What is the area under $f(x)$ lying between a and b ?

Question (i) is answered by looking at chords of $f(x)$. For any two points a and x , the slope of the chord from $f(a)$ to $f(x)$ is

$$s(x) = \frac{f(x) - f(a)}{x - a}.$$

If $s(x)$ has a limit as $x \rightarrow a$, then this is what we regard as the slope of $f(x)$ at a . We call it the *derivative* of $f(x)$, and say that $f(x)$ is *differentiable* at a .

Derivative The derivative of $f(x)$ at a is denoted by $f'(a)$, where

$$f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}. \quad \blacktriangle$$

We also write this as

$$f'(a) = \left. \frac{df}{dx} \right|_{x=a}.$$

In this notation $df/dx = df(x)/dx$ is the function of x that takes the value $f'(a)$ when $x = a$.

If we can differentiate the derivative $f'(x)$, then we obtain the second derivative denoted by $f^{(2)}(x)$. Continuing in this way, the n th derivative of $f(x)$ is $f^{(n)}(x) = \frac{d^n f}{dx^n}$.

For question (ii), let $f(x)$ be a function defined on $[a, b]$. Then the area under the curve $f(x)$ in $[a, b]$ is denoted by

$$\int_a^b f(x)dx,$$

and is called the integral of $f(x)$ from a to b . In general, areas below the x -axis are counted as negative; for a probability density, this case does not arise because density functions are never negative.

The integral is also defined as a limit, but any general statements would take us too far afield. For well-behaved positive functions, you can determine the integral as follows. Plot $f(x)$ on squared graph paper with interval length $1/n$. Let S_n be the number of squares lying entirely between $f(x)$ and the x -axis between a and b . Set $I_n = S_n/n^2$. Then

$$\lim_{n \rightarrow \infty} I_n = \int_a^b f(x)dx.$$

The function $f(x)$ is said to be integrable.

Of course, we almost never obtain integrals by performing such a limit. We almost always use a method that relies on the following, most important, connection between differentiation and integration.

Fundamental Theorem of Calculus

Let $f(x)$ be a continuous function defined on $[a, b]$, and suppose that $f(x)$ is integrable. Define the function $F_a(x)$ by

$$F_a(x) = \int_a^x f(t)dt.$$

Then the derivative of $F_a(x)$ is $f(x)$; formally,

$$F'_a(x) = f(x).$$

This may look like sorcery, but actually it is intuitively obvious. The function $F'_a(x)$ is the slope of $F_a(x)$; that is, it measures the rate at which area is appearing under $f(x)$ as x increases. Now just draw a picture of $f(x)$ to see that extra area is obviously appearing at the rate $f(x)$, so $F'_a(x) = f(x)$. We omit any proof.

Summary of elementary properties

(i) $f'(x) = df/dx = f'$. It follows that

If f is constant, then $f' = 0$.

$$\frac{d}{dx}(cf + g) = cf' + g', \quad \text{for constant } c.$$

$$\frac{d}{dx}(fg) = f'g + fg'.$$

$$\frac{d}{dx}f(g) = f'(g)g'.$$

(ii) $F(x) = \int_{-\infty}^x f(t)dt$. It follows that

If f is constant, then $F(b) - F(a) = \int_a^b f(t)dt \propto b - a$.

If $f < g$, then $\int_a^b f dx < \int_a^b g dx$.

$$\int_a^b (cf + g)dx = c \int_a^b f dx + \int_a^b g dx.$$

$$\int_a^b f'g dx + \int_a^b fg' dx = \int_a^b (fg)' dx = f(b)g(b) - f(a)g(a).$$

(iii) $\log x = \int_1^x (1/y)dy$.

Functions of More Than One Variable

We note briefly that the above ideas can be extended quite routinely to functions of more than one variable. For example, let $f(x, y)$ be a function of x and y . Then

(i) $f(x, y)$ is continuous in x at (a, y) if $\lim_{x \rightarrow a} f(x, y) = f(a, y)$.

(ii) $f(x, y)$ is continuous at (a, b) if $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b)$.

(iii) $f(x, y)$ is differentiable in x at (a, y) if

$$\lim_{x \rightarrow a} \frac{f(x, y) - f(a, y)}{x - a} = f_1(a, y)$$

exists. We denote this limit by $\partial f/\partial x$. That is to say, $\partial f/\partial x$ is the function of x and y that takes the value $f_1(a, y)$ when $x = a$. Other derivatives, such as $\partial f/\partial y$ and $\partial^2 f/\partial x \partial y$, are defined in exactly the same way.

Finally, we note a small extension of the fundamental theorem of calculus, which is used more often than you might expect:

$$\frac{\partial}{\partial x} \int^{g(x)} f(u, y) du = \frac{dg(x)}{dx} f(g(x), y).$$

WORKED EXAMPLES AND EXERCISES

6.10 Example: Gambler's Ruin and First Passages

Let $S_n = a + \sum_{i=1}^n X_i$ be a simple random walk with $a \geq 0$.

- (a) Suppose that $0 \leq S_n \leq K$, and that the walk stops as soon as either $S_n = 0$ or $S_n = K$. (In effect, this is the gambler's ruin problem.) Define $T_{a0} = \min\{n : S_n = 0\}$, and find $\mathbf{E}(s^{T_{a0}}) = F_a(s)$ (say).
- (b) Now suppose that $K = \infty$, and define $T_{a0} = \min\{n : S_n = 0\}$. (In effect, this is the gambler's ruin with an infinitely rich opponent.) Find $\mathbf{E}(s^{T_{10}}) = F_{1,0}(s)$ (say).

Solution (a) Of course, T_{a0} is defective in general because the walk may stop at K . This makes no difference to the fact that, by conditional expectation, for $0 < a < K$,

$$(1) \quad \begin{aligned} F_a(s) &= \mathbf{E}(\mathbf{E}(s^{T_{a0}} | X_1)) = p\mathbf{E}(s^{T_{a0}} | X_1 = 1) + q\mathbf{E}(s^{T_{a0}} | X_1 = -1) \\ &= ps\mathbf{E}(s^{T_{a+1,0}}) + qs\mathbf{E}(s^{T_{a-1,0}}) = psF_{a+1}(s) + qsF_{a-1}(s). \end{aligned}$$

Because the walk stops at 0 or K ,

$$(2) \quad F_K(s) = 0 \quad \text{and} \quad F_0(s) = 1.$$

The difference equation (1) has auxiliary equation $psx^2 - x + qs = 0$, with roots

$$\begin{aligned} \lambda_1(s) &= (1 + (1 - 4pqs^2)^{\frac{1}{2}})/(2ps) \\ \lambda_2(s) &= (1 - (1 - 4pqs^2)^{\frac{1}{2}})/(2ps). \end{aligned}$$

Hence, in the by now familiar way, the solution of (1) that satisfies (2) is

$$(3) \quad F_a(s) = \frac{\lambda_1^K \lambda_2^a - \lambda_1^a \lambda_2^K}{\lambda_1^K - \lambda_2^K}.$$

(b) When the walk is unrestricted it is still the case, by conditional expectation, that

$$(4) \quad F_{1,0}(s) = \mathbf{E}(\mathbf{E}(s^{T_{10}} | X_1)) = ps\mathbf{E}(s^{T_{20}}) + qs.$$

However, using the same argument as we did above for the symmetric random walk yields $\mathbf{E}(s^{T_{20}}) = (\mathbf{E}(s^{T_{10}}))^2$. Substituting in (4) shows that $F_{1,0}(s)$ is either $\lambda_1(s)$ or $\lambda_2(s)$. However, $\lambda_1(s)$ is not the p.g.f. of a nonnegative random variable and so

$$(5) \quad F_{1,0}(s) = \lambda_2(s) = (1 - (1 - 4pqs^2)^{\frac{1}{2}})/(2ps).$$

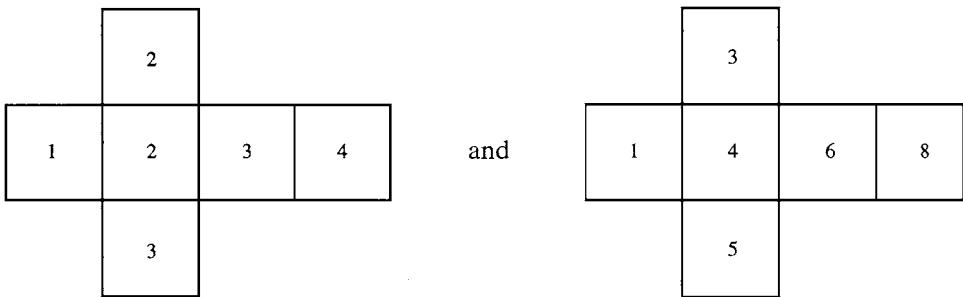
- (6) **Exercise** Let $T_{aK} = \min \{n: S_n = K\}$.
Show that

$$\mathbf{E}(s^{T_{aK}}) = \frac{\lambda_1^a - \lambda_2^a}{\lambda_1^K - \lambda_2^K}.$$

- (7) **Exercise** What is the p.g.f. of the duration of the game in the gambler's ruin problem?
 (8) **Exercise** What is $\mathbf{E}(s^{T_{a0}})$ for the unrestricted random walk?
 (9) **Exercise** For the unrestricted random walk started at $a > 0$, find the probability that the walk ever visits 0 and $\mathbf{E}(T_{a0} | T_{a0} < \infty)$.
 (10) **Exercise** Let S_n be a simple random walk with $S_0 = 0$. Let $T = \min \{n > 0: S_n = 0\}$. Show that $\mathbf{E}(s^T) = 1 - (1 - 4pq s^2)^{\frac{1}{2}}$. What is $\mathbf{E}(T | T < \infty)$?

6.11 Example: "Fair" Pairs of Dice

You have the opportunity to play a game of craps with either "Lucky" Luke or "Fortunate" Fred. Whose dice shall you play with? Luke's two dice are perfectly regular cubes, but the faces bear unorthodox numbers:



Luke explains that, when rolled, the sum of his two dice has the same mass function as the sum of two conventional fair dice; he uses these to ensure that no one can surreptitiously switch to unfair dice.

Fred's two dice are conventionally numbered, but are irregular cubes. Fred explains that these have been cleverly biased so that, when rolled, the sum has the same mass function as two fair dice; their irregular shape ensures that no one can secretly switch to unfair dice.

Assuming you want to play at the usual odds, whose dice should you use? (Sadly, your own dice were confiscated by a casino last week.)

Solution Let X and Y be the scores of two fair dice. The p.g.f. of their sum is

$$\begin{aligned} \mathbf{E}(s^{X+Y}) &= \mathbf{E}(s^X)\mathbf{E}(s^Y) = \left(\frac{1}{6}(s + s^2 + s^3 + s^4 + s^5 + s^6)\right)^2 \\ &= \frac{s^2}{36}(1 + 2s + 3s^2 + 4s^3 + 5s^4 + 6s^5 + 5s^6 + 4s^7 + 3s^8 + 2s^9 + s^{10}) \\ &= G(s) \quad (\text{say}). \end{aligned}$$

Now the sum of Luke's dice $L_1 + L_2$ has p.g.f.

$$\mathbf{E}(s^{L_1+L_2}) = \frac{1}{6}(s + 2s^2 + 2s^3 + s^4) \frac{1}{6}(s + s^3 + s^4 + s^5 + s^6 + s^8) = G(s)$$

on multiplying out the brackets. So Luke's claim is correct. However, $G(s)$ can be factorized as

$$36G(s) = s^2(1+s)^2(1-s+s^2)^2(1+s+s^2)^2,$$

where $1+s+s^2$, $1-s+s^2$ are irreducible, having complex roots. Hence, there are only two possibilities for the generating functions of Fred's dice:

either

$$(i) \quad \mathbf{E}(s^{F_1}) = \frac{1}{2}s(1+s)(1-s+s^2)^2$$

and

$$\mathbf{E}(s^{F_2}) = \frac{s}{18}(1+s)(1+s+s^2)^2;$$

or (ii) the dice are fair:

$$\begin{aligned} \mathbf{E}(s^{F_1}) &= \mathbf{E}(s^{F_2}) = \frac{1}{6}s(1+s)(1+s+s^2)(1-s+s^2) \\ &= \frac{1}{6}(s + s^2 + s^3 + s^4 + s^5 + s^6). \end{aligned}$$

However, $(1+s)(1-s+s^2)^2 = 1-s+s^2+s^3-s^4+s^5$ and the negative coefficients ensure that $\frac{1}{2}s(1+s)(1-s+s^2)^2$ is not a p.g.f. The only remaining possibility is that the dice are fair, which palpably they are not.

This shows that the sum of two biased dice cannot have the same mass function as the sum of two fair dice. Thus, Fred's claim is incorrect; his dice are as crooked as yours probably were.

You should play with Luke's dice.

- (1) **Exercise** You have two fair tetrahedral dice whose faces are numbered conventionally 1, 2, 3, 4. Show how to renumber the faces so that the distribution of the sum is unchanged.
- (2) **Exercise** Yet another regular Platonic solid is the dodecahedron with 12 pentagonal faces.
 - (a) Write down the generating function of the sum of two fair dodecahedra with faces numbered 1 to 12 inclusive.
 - (b) Show that two such dodecahedral dice can be biased in such a way that their sum has the same distribution as the sum of the fair dice.

Hint: Let

$$\begin{aligned} f(x) &= x + x^{12} + (2 - \sqrt{3})(x^2 + x^{11}) + (5 - 2\sqrt{3})(x^3 + x^{10}) \\ &\quad + (7 - 4\sqrt{3})(x^4 + x^9) + (10 - 5\sqrt{3})(x^5 + x^8) + (11 - 6\sqrt{3})(x^3 + x^7) \end{aligned}$$

and $g(x) = (x + x^4 + x^7 + x^{10})(1 + \sqrt{3}x + x^2)$. Consider $f(x)g(x)$.

- (3) **Exercise** Show that it is not possible to weight two conventional dice in such a way that the sum of the numbers shown is equally likely to take any value between 2 and 12 inclusive.
- (4) **Exercise** Is it possible to re-number two fair dice so that their sum is equally likely to take any value between 2 and 12 inclusive?
- (5) **Exercise** Can the sum of three biased dice have the same mass function as the sum of three fair dice?

Remark Some results of this example were recorded by SG Landry, LC Robertson and RM Shortt in the *American Mathematical Monthly*, 1988.

6.12 Example: Branching Process

Let Z_n be the size at time n of the ordinary branching process defined in Example 6.3.16. Thus the r th individual in the n th generation (that is, at time n) is replaced by a family of size $X(r, n + 1)$, where the $X(r, n + 1)$ are independent and identically distributed, with mean μ , variance σ^2 , and cumulant generating function

$$(1) \quad \kappa(t) = \log(\mathbf{E}[\exp(tX(1, 1))]).$$

Show that

$$(2) \quad \mathbf{E}(Z_n) = \mu^n.$$

Show also that

$$(3) \quad \text{var}(Z_n) = \text{var}(Z_1)(\mathbf{E}(Z_1))^{n-1} + (\mathbf{E}(Z_1))^2 \text{var}(Z_{n-1})$$

and hence find an expression for $\text{var}(Z_n)$ in terms of μ and σ , when $\mu \neq 1$.

Solution First, recall the basic identity of branching processes: namely, given Z_{n-1} ,

$$(4) \quad Z_n = \sum_{r=1}^{Z_{n-1}} X(r, n).$$

Now let the cumulant generating function of Z_n be $\kappa_n(t)$. Then by conditional expectation,

$$(5) \quad \begin{aligned} \kappa_n(t) &= \log(\mathbf{E}(e^{tZ_n})) = \log(\mathbf{E}(\mathbf{E}(e^{tZ_n} | Z_{n-1}))) \\ &= \log\left(\mathbf{E}\left(\mathbf{E}\exp\left[t\sum_{r=1}^{Z_{n-1}} X(r, n)\right] | Z_{n-1}\right)\right) \quad \text{by (4)} \\ &= \log(\mathbf{E}[(\mathbf{E}(e^{tX(1,1)})^{Z_{n-1}})]) = \log(\mathbf{E}[(e^{\kappa(t)}]^{Z_{n-1}})) \\ &= \kappa_{n-1}(\kappa(t)). \end{aligned}$$

Now expanding $\kappa_{n-1}(\kappa(t))$ using (6.4.12) gives

$$(6) \quad \begin{aligned} \kappa_{n-1}(\kappa(t)) &= \kappa(t)\mathbf{E}(Z_{n-1}) + \frac{1}{2}(\kappa(t))^2 \text{var}(Z_{n-1}) + \dots \\ &= \mu t \mathbf{E}(Z_{n-1}) + \frac{1}{2} \sigma^2 \mathbf{E}(Z_{n-1}) t^2 + \frac{1}{2} \mu^2 \text{var}(Z_{n-1}) t^2 + O(t^3) \end{aligned}$$

on expanding $\kappa(t)$ using (6.4.12). Hence, equating coefficients of t and t^2 in (5) and (6) now yields

$$(7) \quad \mathbf{E}(Z_n) = \mu \mathbf{E}(Z_{n-1})$$

and

$$(8) \quad \text{var}(Z_n) = \sigma^2 \mu^{n-1} + \mu^2 \text{var}(Z_{n-1}).$$

Iterating (7) gives (2), and equation (8) is just (3), as required. To solve the difference equation (8), we note first that $A\mu^{2n}$ is a solution of the reduced equation

$$\text{var}(Z_n) = \mu^2 \text{var}(Z_{n-1}).$$

By inspection, a particular solution of (8), when $\mu \neq 1$, is given by $\sigma^2 \mu^{n-1} / (1 - \mu)$. Imposing the initial condition $\text{var}(Z_1) = \sigma^2$ now shows that when $\mu \neq 1$, (8) has solution

$$(9) \quad \text{var}(Z_n) = \sigma^2 \mu^{n-1} \frac{1 - \mu^n}{1 - \mu}.$$

(10) **Exercise** Find $\text{var}(Z_n)$ when $\mu = 1$.

(11) **Exercise** Show that for $n > m$, $\mathbf{E}(Z_n Z_m) = \mu^{n-m} \mathbf{E}(Z_m^2)$.

Deduce that when $\mu = 1$,

$$\rho(Z_m, Z_n) = \left(\frac{m}{n}\right)^{\frac{1}{2}}.$$

(12) **Exercise** Find an expression for $\rho(Z_m, Z_n)$ when $\mu \neq 1$, and deduce that for $\mu > 1$ as $n, m \rightarrow \infty$, with $n - m$ held fixed, $\rho(Z_m, Z_n) \rightarrow 1$.

(13) **Exercise** If r is such that $r = \mathbf{E}(r^{Z_1})$, show that $\mathbf{E}(r^{Z_{n+1}} | Z_n) = r^{Z_n}$. What is r ?

6.13 Example: Geometric Branching

Let $(Z_n; n \geq 0)$ be an ordinary branching process with $Z_0 = 1$, and suppose that

$$(1) \quad \mathbf{E}(s^{Z_1}) = \frac{1 - p}{1 - ps}.$$

(a) Show that, for $p \neq \frac{1}{2}$,

$$(2) \quad \mathbf{E}(s^{Z_n}) = \frac{\rho^n - 1 - \rho s(\rho^{n-1} - 1)}{\rho^{n+1} - 1 - \rho s(\rho^n - 1)},$$

where $\rho = p/(1 - p)$.

(b) Now let $(Z_n^*; n \geq 0)$ be an ordinary branching process with $Z_0 = 1$, family size distribution given by (1), and such that at time n , for all $n \geq 1$, one new particle is added to the population independently of Z_n .

Show that for $p < \frac{1}{2}$

$$(3) \quad \lim_{n \rightarrow \infty} \mathbf{E}(s^{Z_n} | Z_n > 0) = \lim_{n \rightarrow \infty} \mathbf{E}(s^{Z_n^*}) = \frac{s(1 - 2p)}{1 - p(1 + s)}.$$

Solution (a) As usual, let $\mathbf{E}(s^{Z_n}) = G_n(s)$. We establish (2) by induction. Assuming that (2) holds for n , it follows from Example 6.3.16 that

$$\begin{aligned}
 G_{n+1}(s) &= G_n\left(\frac{1-p}{1-ps}\right) \\
 &= \frac{(\rho^n - 1)(1-ps) - p(\rho^{n-1} - 1)}{(\rho^{n+1} - 1)(1-ps) - p(\rho^n - 1)} \quad \text{by the induction hypothesis} \\
 (4) \quad &= \frac{\rho^{n+1} - 1 - \rho s(\rho^n - 1)}{\rho^{n+2} - 1 - \rho s(\rho^{n+1} - 1)}.
 \end{aligned}$$

Because (2) is true for $n = 1$, by (1), the result does follow by induction.

(b) Let $(\hat{Z}_n; n \geq 1)$ be a collection of independent random variables such that \hat{Z}_n has the same distribution as Z_n . Now Z_n^* is the sum of the descendants of the initial individual, and the descendants of the fresh individual added at $n = 1$, and those of the next added at $n = 2$, and so on. That is to say Z_n^* has the same distribution as $1 + \hat{Z}_1 + \dots + \hat{Z}_n$. Hence,

$$\begin{aligned}
 (5) \quad \mathbf{E}(s^{Z_n^*}) &= s \prod_{r=1}^n G_r(s) = \frac{s(\rho - 1)}{\rho^{n+1} - 1 - \rho s(\rho^n - 1)} \quad \text{by successive cancellation,} \\
 &\rightarrow \frac{s(\rho - 1)}{\rho s - 1} \quad \text{as } n \rightarrow \infty \\
 &= \frac{s(1 - 2p)}{1 - p(1 + s)}.
 \end{aligned}$$

This is the generating function of a random variable with mass function

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

For the other half of (3), we require the conditional generating function $\mathbf{E}(s^{Z_n} | Z_n > 0)$. Because $\mathbf{P}(Z_n = 0) = G_n(0)$, this is given by

$$(6) \quad \mathbf{E}(s^{Z_n} | Z_n > 0) = \frac{G_n(s) - G_n(0)}{1 - G_n(0)}.$$

Substituting for $G_n(\cdot)$ from (2), we find

$$\begin{aligned}
 \mathbf{E}(s^{Z_n} | Z_n > 0) &= \left(\frac{\rho^n - 1 - s\rho^n + \rho s}{\rho^{n+1} - 1 - s\rho^{n+1} + \rho s} - \frac{\rho^n - 1}{\rho^{n+1} - 1}\right) \bigg/ \left(1 - \frac{\rho^n - 1}{\rho^{n+1} - 1}\right) \\
 &\rightarrow \frac{s(1 - \rho)}{1 - \rho s} \quad \text{as } n \rightarrow \infty,
 \end{aligned}$$

as required.

(7) **Exercise** Let Z_n be an ordinary branching process with family size mass function $\mathbf{P}(X = k) = 2^{-(k+1)}$; $k \geq 0$. Show that

$$G_n(s) = \frac{n - (n-1)s}{n + 1 - ns}; \quad n \geq 0.$$

- (8) **Exercise (7) Continued** Show that in this case ($p = \frac{1}{2}$), we have

$$\mathbf{E} \left(\frac{Z_n}{n} \mid Z_n > 0 \right) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

- (9) **Exercise** (a) Let X be any nonnegative random variable such that $\mathbf{E}(X) = 1$. Show that $\mathbf{E}(X \mid X > 0) \leq \mathbf{E}(X^2)$.
 (b) Deduce that if $p > \frac{1}{2}$, $\mathbf{E}(Z_n \rho^{-n} \mid Z_n > 0) < 2p/(p - q)$.
- (10) **Exercise** When $p < q$, find $\lim_{n \rightarrow \infty} \mathbf{E}(s^{Z_n} t^{Z_{n+m}})$.

6.14 Example: Waring's Theorem: Occupancy Problems

Let $(A_i; 1 \leq i \leq n)$ be a collection of events (not necessarily independent). Let X be the number of these events that occur, and set $p_m = \mathbf{P}(X = m)$, $q_m = \mathbf{P}(X \geq m)$, and

$$s_m = \sum_{i_1 < \dots < i_m} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_m}).$$

Show that

$$(a) \quad s_m = \mathbf{E} \binom{X}{m} = \sum_{i=m}^n \binom{i}{m} p_i;$$

$$(b) \quad p_m = \sum_{i=m}^n (-1)^{i-m} \binom{i}{m} s_i;$$

and (c)
$$q_m = \sum_{i=m}^n (-1)^{i-m} \binom{i-1}{m-1} s_i.$$

Solution (a) Recall that $I(A)$ is the indicator of the event A . Because

$$\mathbf{P}(A_{i_1} \cap \dots \cap A_{i_m}) = \mathbf{E}(I(A_{i_1}) \dots I(A_{i_m}))$$

it follows that the sum s_m is just the expected number of distinct sets of size m (of the A_i) that occur. But, given $X = i$, exactly $\binom{i}{m}$ such distinct sets of size m occur. Hence, by conditional expectation,

$$(1) \quad s_m = \sum_{i=m}^n \binom{i}{m} \mathbf{P}(X = i) = \mathbf{E} \binom{X}{m}.$$

(b) Now define generating functions $G_s(z) = \sum_{m=0}^n z^m s_m$ and $G_q(z) = \sum_{m=0}^n z^m q_m$. By (1),

$$(2) \quad \begin{aligned} G_s(z-1) &= \sum_{m=0}^n (z-1)^m s_m = \sum_{m=0}^n (z-1)^m \mathbf{E} \binom{X}{m} = \mathbf{E} \left(\sum_{m=0}^X (z-1)^m \binom{X}{m} \right) \\ &= \mathbf{E}((1+z-1)^X) = \mathbf{E}(z^X) = \sum_{m=0}^n p_m z^m. \end{aligned}$$

Now equating coefficients of z_m proves (b).

(c) By Theorem 3.6.7 or Theorem 6.1.5, we have

$$G_q(z) = \frac{1 - zG_X(z)}{1 - z}.$$

Hence, by (2),

$$(3) \quad \frac{G_q(z) - 1}{z} = \frac{G_s(z - 1) - 1}{z - 1}.$$

Equating coefficients of z^m yields (c).

(4) **Exercise** Show that

$$s_m = \sum_{i=m}^n \binom{i-1}{m-1} q_i.$$

(5) **Exercise** Let $t_m = \sum_{i_1 < \dots < i_m} \mathbf{P}(A_{i_1} \cup \dots \cup A_{i_m})$. Find expressions for s_m , p_m , and q_m in terms of $(t_i; 1 \leq i \leq n)$.

(6) **Exercise** If r balls are placed at random in n cells, so that each ball is independently equally likely to arrive in any cell, find the probability that exactly c cells are each occupied by exactly i balls.

6.15 Example: Bernoulli Patterns and Runs

A coin is tossed repeatedly, heads appearing with probability $p (= 1 - q)$ on each toss.

(a) Let X be the number of tosses required until the first occasion when successive tosses show HTH . Show that

$$(1) \quad \mathbf{E}(s^X) = \frac{p^2qs^3}{1 - s + pqs^2 - pq^2s^3}.$$

(b) Let Y be the number of tosses required until the first occasion when three successive tosses show HTH or THT . Show that

$$(2) \quad \mathbf{E}(s^Y) = \frac{pqs^3(1 - 2pqs + pqs^2)}{1 - s + pqs^2 - p^2q^2s^4}.$$

Solution (a) Consider the event that HTH does not appear in the first n tosses, and it does then appear as a result of the next three. The probability of this is $\mathbf{P}(X > n)p^2q$.

This event is the union of the following two disjoint events:

- (i) The last two tosses of the n were HT , in which case $X = n + 1$.
- (ii) The last two tosses of the n were not HT , in which case $X = n + 3$.

The probabilities of these two events are $\mathbf{P}(X = n + 1)pq$ and $\mathbf{P}(X = n + 3)$, respectively. Hence,

$$(3) \quad \mathbf{P}(X > n)p^2q = \mathbf{P}(X = n + 1)pq + \mathbf{P}(X = n + 3); \quad n \geq 0.$$

Multiplying (3) by s^{n+3} , summing over n , and recalling (6.1.7), yields

$$(4) \quad \frac{1 - \mathbf{E}(s^X)}{1 - s} p^2qs^3 = pqs^2\mathbf{E}(s^X) + \mathbf{E}(s^X).$$

The required result (1) follows.

(b) First, consider the event that neither *HTH* nor *THT* have appeared in the first n tosses and *HTH* then appears. This event is the union of three disjoint events:

- (i) The last two tosses of the n were *HT*, so $Y = n + 1$, with the appearance of *HTH*;
- (ii) The last two tosses of the n were *TT*, so $Y = n + 2$, with the appearance of *THT*;
- (iii) Otherwise, $Y = n + 3$, with the appearance of *HTH*.

Let $f^1(n)$ denote the probability that $Y = n$ with the occurrence of *HTH*, and $f^2(n)$ the probability that $Y = n$ with the occurrence of *THT*. Then from the above, we have

$$(5) \quad \mathbf{P}(Y > n)p^2q = f^1(n+1)pq + f^2(n+2)p + f^1(n+3).$$

Hence, multiplying by s^{n+3} and summing over n ,

$$(6) \quad \frac{1 - \mathbf{E}(s^Y)}{1 - s} p^2qs^3 = pqs^2G^1(s) + spG^2(s) + G^1(s),$$

where $G^i(s) = \sum_n s^n f^i(n)$; $i = 1, 2$.

Second, consider the event that neither *HTH* or *THT* have appeared in the first n tosses and *THT* then occurs. Arguing as in the first case yields

$$(7) \quad \mathbf{P}(Y > n)pq^2 = f^2(n+1)pq + f^1(n+2)q + f^1(n+3).$$

Hence,

$$(8) \quad \frac{1 - \mathbf{E}(s^Y)}{1 - s} pq^2s^3 = pqs^2G^2(s) + qsG^1(s) + G^2(s).$$

Now we also have

$$(9) \quad \mathbf{E}(s^Y) = G^1(s) + G^2(s)$$

so solving (6), (8), and (9) for $\mathbf{E}(s^Y)$ yields (2) as required.

- (10) **Exercise** When $p = \frac{1}{2}$, find $\mathbf{E}(X)$ for all possible triples of the form *HHH*, *HHT*, etc. Comment on your results.
- (11) **Exercise** Show that $\mathbf{E}(X) = 1/p + 1/(qp^2)$.
- (12) **Exercise** Show that $\mathbf{E}(Y) = (1 + pq + p^2q^2)/(pq(1 - pq))$.
- (13) **Exercise** Let Z be the number of tosses required for the first occurrence of *HHH*. Find $\mathbf{E}(s^Z)$ and show that

$$\mathbf{E}(Z) = \frac{1}{p} + \frac{1}{p^2} + \frac{1}{p^3}.$$

- (14) **Exercise** Let W be the number of tosses required for the first appearance of either *HHH* or *TTT*. Show that

$$\mathbf{E}(s^W) = \frac{s^3(p^3 + q^3 + qp(p^2 + q^2)s + p^2q^2s^2)}{1 - pq s^2 - pq s^3 - p^2q^2s^4}.$$

- (15) **Exercise** Let V be the number of tosses required for the first appearance of either r consecutive heads or ρ consecutive tails. Find $\mathbf{E}(s^V)$, and show that

$$\mathbf{E}(V) = \left(\frac{p^r(1-p)}{1-p^r} + \frac{q^\rho(1-q)}{1-q^\rho} \right)^{-1}.$$

- (16) **Exercise** Find the expected number of tosses required for the first appearance of *HTHTH*.

6.16 Example: Waiting for Unusual Light Bulbs

The light bulbs in the sequence illuminating your room have independent and identically distributed lifetimes, so the replacement times form an ordinary renewal process (as defined in Example 6.7.2). Suppose the lifetimes are $(X_i; i \geq 1)$ with common mass function $f(k)$.

A bulb is called “unusual” if its life is shorter than a or longer than b , where $a \leq b$. Let T be the time at which a bulb is first identified as being an unusual bulb. Show that (for integers a and b),

$$(1) \quad \mathbf{E}(s^T) = \frac{\mathbf{E}(I_a s^{X_1}) + s^b \mathbf{E}(I_b)}{1 - \mathbf{E}(s^{X_1} I_a^c I_b^c)}.$$

where I_a and I_b are the indicators of the events $\{X_1 < a\}$ and $\{X_1 > b\}$ respectively. That is,

$$I_a = I\{X_1 < a\}, \quad \text{with } I_a^c = 1 - I_a,$$

and

$$I_b = I\{X_1 > b\}, \quad \text{with } I_b^c = 1 - I_b.$$

Solution Because $I_a + I_b + I_a^c I_b^c = 1$, we can write

$$(2) \quad \mathbf{E}(s^T) = \mathbf{E}(I_a s^T) + \mathbf{E}(I_b s^T) + \mathbf{E}(I_a^c I_b^c s^T).$$

Now on the event I_a , $X_1 = T$ because the first bulb failed before a and was identified as unusual at X_1 . So

$$(3) \quad \mathbf{E}(I_a s^T) = \mathbf{E}(I_a s^{X_1}).$$

On the event $I_a^c I_b^c$, the process regenerates at the first replacement $X_1 \in [a, b]$, and so $T = X_1 + T'$, where T' is independent of X_1 and has the same distribution as T . So

$$(4) \quad \mathbf{E}(I_a^c I_b^c s^T) = \mathbf{E}(s^{X_1}) \mathbf{E}(I_a^c I_b^c s^{X_1}).$$

Finally on the event I_b , the first light bulb is identified as unusual when it survives beyond b , so

$$(5) \quad \mathbf{E}(I_b s^T) = s^b \mathbf{E}(I_b).$$

Substituting (3), (4), and (5) into (2), gives (1).

(6) **Exercise** Find the expected time until a light bulb has a lifetime shorter than a . Evaluate this when $f_X(k) = qp^{k-1}$, $k \geq 1$.

(7) **Exercise: Crossing the Road** Successive cars pass at instants X_i seconds apart ($i \geq 1$). You require b seconds to cross the road. If the random variables $(X_i; i \geq 1)$ are independent and identically distributed, find your expected waiting time until you cross. Evaluate this when $f_X(k) = qp^{k-1}$; $k \geq 1$.

(8) **Exercise** Let L be the time until a light bulb has lasted longer than r . Show that

$$\mathbf{E}(s^L) = \frac{s^r \mathbf{P}(X > r)}{1 - \sum_{k=1}^r s^k \mathbf{P}(X = k)}.$$

- (9) **Exercise** A biased coin is tossed repeatedly; on each toss, it shows a head with probability $p (= 1 - q)$. Let W be the number of tosses until the first occasion when r consecutive tosses have shown heads. Show that

$$\mathbf{E}(s^W) = \frac{(1 - ps)p^r s^r}{1 - s + qp^r s^{r+1}}.$$

- (10) **Exercise** In n tosses of a biased coin, let L_n be the length of the longest run of heads, and set $\pi_{n,r} = \mathbf{P}(L_n < r)$. Show that

$$1 + \sum_{n=1}^{\infty} s^n \pi_{n,r} = \frac{1 - p^r s^r}{1 - s + qp^r s^{r+1}}.$$

6.17 Example: Martingales for Branching

Let $G_n(s)$ be the probability generating function of the size Z_n of the n th generation of a branching process (as defined in example 6.3.17), where $Z_0 = 1$ and $\text{var } Z_1 > 0$. Let H_n be the inverse function of the function G_n , and show that $M_n = (H_n(s))^{Z_n}$ defines a martingale with respect to $(Z_n; n \geq 0)$.

Solution Because $\text{var } Z_1 > 0$, the function $G(s)$ is strictly increasing on $[0, 1)$. Hence, likewise, so are all the functions $G_n(s)$. By definition,

$$(1) \quad G_n(H_n(s)) = s,$$

and

$$s = G_{n+1}(H_{n+1}(s)) = G_n(G(H_{n+1}(s)))$$

by (6.3.18). Hence, by (1), because $H_n(s)$ is unique,

$$(2) \quad G(H_{n+1}(s)) = H_n(s).$$

Finally, using (6.3.18) again,

$$\mathbf{E}[(H_{n+1}(s))^{Z_{n+1}} | Z_0, \dots, Z_n] = [G(H_{n+1}(s))]^{Z_n} = [H_n(s)]^{Z_n} \quad \text{by (2)}.$$

Trivially, $\mathbf{E}M_n = 1$; it follows that M_n is a martingale.

- (3) **Exercise** Show that η^{Z_n} is a martingale where η is the extinction probability defined in (6.3.20).
 (4) **Exercise** If $\mathbf{E}Z_1 = \mu$, show that $Z_n \mu^{-n}$ is a martingale.
 (5) **Exercise** Let Z_n be the size of the n th generation of the branching process in which the n th generation is augmented by a random number I_n of immigrants who are indistinguishable from the other members of the population, and such that the I_n are independent and identically distributed, and independent of the process up to time n . If $\mathbf{E}I_n = m$, and the expected family size is not 1, show that

$$M_n = \mu^{-n} \left(Z_n - m \frac{1 - \mu^n}{1 - \mu} \right)$$

is a martingale.

6.18 Example: Wald's Identity

Let $(X_n; n \geq 1)$ be independent and identically distributed with $M(t) = \mathbf{E}e^{tX_1}$. Define $S_n = \sum_{r=1}^n X_r$,

and $Y_n = \exp(tS_n)M(t)^{-n}$, $n \geq 1$, with $Y_0 = 1$.

Suppose that T is a stopping time for Y_n , with $\mathbf{E}T < \infty$, and $|S_n| \leq K < \infty$ for $n < T$. Show that, whenever $1 \leq M(t) < \infty$, Y_n is a martingale and

$$\mathbf{E}[\exp(tS_T)M(t)^{-T}] = 1.$$

Solution From the independence of the X_n , it easily follows that Y_n is a martingale. Now

$$\begin{aligned} \mathbf{E}(|Y_{n+1} - Y_n| | Y_0, \dots, Y_n) &= Y_n \mathbf{E} \left(\left| \frac{e^{tX_{n+1}}}{M(t)} - 1 \right| \right) \\ &\leq Y_n \mathbf{E}(e^{tX_{n+1}} + M(t)) / M(t) = 2Y_n. \end{aligned}$$

Hence, for $n < T$,

$$\begin{aligned} \mathbf{E}(|Y_{n+1} - Y_n| | Y_0, \dots, Y_n) &\leq 2Y_n \leq \frac{2e^{t|K}}{M(t)^n} \\ &\leq 2e^{t|K}. \end{aligned}$$

Because $\mathbf{E}T \leq K < \infty$, we can use the final part of the optional stopping theorem 5.7.14 to obtain $\mathbf{E}Y_T = 1$, which is the required result.

- (1) **Exercise** Let $\text{var } X_1 > 0$, and let T be the smallest n such that either $S_n \leq -a < 0$ or $S_n \geq b > 0$. Show that $\mathbf{E}T \leq K < \infty$.
- (2) **Exercise** Assume there is some $t \neq 0$ such that $M(t) = 1$, and let T be defined as in exercise (1). Show that, approximately,

$$\mathbf{P}(S_T \leq -a) \simeq \frac{e^{tb} - 1}{e^{tb} - e^{-ta}}, \quad \text{and} \quad \mathbf{P}(S_T \geq b) \simeq \frac{1 - e^{-ta}}{e^{tb} - e^{-ta}}.$$

Deduce also that $\mathbf{P}(S_T \leq -a) \leq e^{-at}$, and $\mathbf{P}(S_T \geq b) \leq e^{-bt}$.

- (3) **Exercise** By differentiating $\mathbf{E}(Y_n | Y_0, \dots, Y_{n-1}) = Y_{n-1}$ for t , and setting $t = 0$, show that the following are martingales. (You may assume that it is justified to interchange the expectation and differentiation.)
 - (a) $S_n - n\mathbf{E}X_1$
 - (b) $(S_n - n\mathbf{E}X_1)^2 - n\text{var } X_1$
 - (c) $(S_n - n\mathbf{E}X_1)^3 - 3(S_n - n\mathbf{E}X_1)\text{var } X_1 - n\mathbf{E}(X_1 - \mathbf{E}X_1)^3$
 If you have a full pen and lots of paper, you can find as many more such martingales as you please.
- (4) **Exercise** Let S_n be a simple random walk with $\mathbf{P}(X_1 = 1) = p = 1 - q = 1 - \mathbf{P}(X_1 = -1)$. Use Wald's identity to show that, when a and b are integers,

$$\mathbf{E}S^T = \frac{\lambda_1^a \lambda_2^a (\lambda_1^b - \lambda_1^b) + \lambda_1^a - \lambda_2^a}{\lambda_1^{a+b} - \lambda_2^{a+b}},$$

where

$$\lambda_{1,2} = \frac{1 \pm (1 - 4pq s^2)^{1/2}}{2ps}.$$

6.19 Example: Total Population in Branching

Let X_n be an ordinary branching process such that $X_0 = 1$, $\mathbf{E}X_1 = \mu$, $\text{var } X_1 = \sigma^2$, and $\mathbf{E}s^{X_1} = G(s)$. If $Y_n = X_0 + X_1 + \cdots + X_n$ and

$$Q_n(s) = \mathbf{E}s^{Y_n}, \quad 0 \leq s \leq 1,$$

show that

$$(1) \quad Q_{n+1}(s) = sG(Q_n(s)).$$

Solution Note that each member of the first generation X_1 gives rise to an independent copy of the branching process. Conditional on X_1 , we may therefore write

$$X_1 + \cdots + X_n = \tilde{Y}_{n-1}^{(1)} + \tilde{Y}_{n-1}^{(2)} + \cdots + \tilde{Y}_{n-1}^{(X_1)},$$

where $\tilde{Y}_{n-1}^{(i)}$ has the same distribution as Y_{n-1} , and has generating function $Q_{n-1}(s)$. Hence, finally,

$$\begin{aligned} \mathbf{E}s^{X_0+X_1+\cdots+X_n} &= s\mathbf{E}(\mathbf{E}(s^{X_1+\cdots+X_n} | X_1)) \\ &= s\mathbf{E}((Q_{n-1})^{X_1}) \\ &= sG(Q_{n-1}(s)). \end{aligned}$$

(2) **Exercise** Deduce that if $Y = \sum_{n=0}^{\infty} X_n$, then $Q(s) = \mathbf{E}s^Y$ satisfies

$$Q(s) = sG(Q(s)), \quad 0 \leq s \leq 1,$$

where $s^\infty \equiv 0$. If $\mu < 1$, show that

(a) $Q(1) = 1$.

(b) $\mathbf{E}Y = (1 - \mu)^{-1}$.

(c) $\text{var } Y = \sigma^2 / (1 - \mu)^3$.

(3) **Exercise** Find $Q(s)$ in the special case when $G(s) = \frac{p}{1-qs}$, $p + q = 1$. Discuss how $Q(s)$ behaves in the two cases $p < q$ and $q < p$.

(4) **Exercise** Suppose that $G(s) = p/(1 - qs)$, $p + q = 1$. Set $Q_n(s) = y_n(s)/x_n(s)$ in (1) to find that $x_n(s)$ satisfies

$$x_n(s) = x_{n-1}(s) - spqx_{n-2}(s),$$

with $x_0 = 1$ and $x_1 = 1 - qs$. Deduce that

$$Q_n(s) = 2ps \frac{(\lambda - 2qs)\lambda^{n-1} - (\mu - 2qs)\mu^{n-1}}{(\lambda - 2qs)\lambda^n - (\mu - 2qs)\mu^n},$$

where $\lambda = 1 + \sqrt{1 - 4spq}$, $\mu = 1 - \sqrt{1 - 4spq}$.

PROBLEMS

- 1 Let $G(s) = \sum_0^\infty f_X(k)s^k$, where $f_X(k) = \mathbf{P}(X = k); k \geq 0$. Show that:
- (a) $\sum_0^\infty \mathbf{P}(X < k)s^k = sG(s)/(1 - s)$.
- (b) $\sum_0^\infty \mathbf{P}(X \geq k)s^k = (1 - sG(s))/(1 - s)$.
- 2 Find the probability generating function of each of the following distributions and indicate where it exists.
- (a) $f(k) = \frac{1}{n}; 1 \leq k \leq n$.
- (b) $f(k) = \frac{1}{2n + 1}; -n \leq k \leq +n$.
- (c) $f(k) = \frac{1}{k(k + 1)}; 1 \leq k$.
- (d) $f(k) = \begin{cases} \frac{1}{2k(k + 1)} & \text{for } k \geq 1 \\ \frac{1}{2k(k - 1)} & \text{for } k \leq -1. \end{cases}$
- (e) $f(k) = \frac{1 - c}{1 + c}c^{|k|}; k \in \mathbb{Z}, 0 < c < 1$.
- 3 Which of the following are probability generating functions, and when?
- (a) $\exp(-\lambda(1 - G_X(s)))$, where $\lambda > 0$, and $G_X(s)$ is a p.g.f.
- (b) $\sin\left(\frac{\pi s}{2}\right)$
- (c) $\left(\frac{q}{1 - ps}\right)^r$
- (d) $(q + ps)^r$
- (e) $1 - (1 - s^2)^{\frac{1}{2}}$
- (f) $\alpha \log(1 + \beta s)$
- 4 If the random variable X has p.g.f. $G(s)$, show that for constants a and b the random variable $aX + b$ has p.g.f. $s^b G(s^a)$. For what values of s is this defined?
- 5 Let X have p.g.f. $G(s)$. Describe a random variable Y , which has p.g.f. $G_Y(s) = G(s)(2 - G(s))^{-1}$. For what values of s is this defined?
- 6 A loaded die may show different faces with different probabilities. Show that it is not possible to load two traditional cubic dice in such a way that the sum of their scores is uniformly distributed on $\{2, 3, \dots, 12\}$.
- 7 The three pairs of opposite faces of a fair die show 1, 2, and 3, respectively. The two faces of a fair coin show 1 and 2, respectively.
- (a) Find the distribution of the sum of their scores when tossed together.
- (b) Is it possible to weight the die in such a way that the sum of the scores is uniform on $\{2, 3, 4, 5\}$?
- 8 Let X have p.g.f. $G(s)$, and let E be the event that X is even. Show that

$$\mathbf{E}(s^X | E) = \frac{G(s) + G(-s)}{G(1) + G(-1)}.$$

- 9 Define the probability generating function of an integer valued random variable X , and show how it may be used to obtain the mean μ_X , variance σ_X^2 , and third moment about the mean γ_X .

- (a) Let $Y = \sum_{i=1}^N X_i$, where the X_i are independent integer valued random variables identically distributed as X . Let $\mu_X = 0$, and let N be an integer valued random variable distributed independently of the X_i . Show that $\sigma_Y^2 = \mu_N \sigma_X^2$, and $\gamma_Y = \mu_N \gamma_X$.
- (b) Find σ_Y^2 when $\mu_X \neq 0$.
- 10** An unfair coin is tossed n times, each outcome is independent of all the others, and on each toss a head is shown with probability p . The total number of heads shown is X . Use the probability generating function of X to find:
- (a) The mean and variance of X .
- (b) The probability that X is even.
- (c) The probability that X is divisible by 3.
- 11** Let the nonnegative random variable X have p.g.f. $G_X(s)$. Show that

$$G(s) = \frac{1}{\mathbf{E}(X)} \cdot \frac{1 - G_X(s)}{1 - s}$$

is the p.g.f. of a nonnegative random variable Y . When is $G(s) = G_X(s)$?

- 12** Let $G_1(s)$ and $G_2(s)$ be probability generating functions, and suppose that $0 \leq \lambda \leq 1$. Show that $\lambda G_1 + (1 - \lambda)G_2$ is a p.g.f., and interpret this result.
- 13** In a multiple-choice examination, a student chooses between one true and one false answer to each question. Assume the student answers at random, and let N be the number of such answers until she first answers two successive questions correctly. Show that $\mathbf{E}(s^N) = s^2(4 - 2s - s^2)^{-1}$. Hence, find $\mathbf{E}(N)$ and $\mathbf{P}(N = k)$. Now find $\mathbf{E}(N)$ directly.
- 14** A number X of objects are ranked in order of beauty (with no ties). You pick one at random with equal probability of picking any.
- (a) If $X - 1$ has a Poisson distribution with parameter λ , show that the p.g.f. of the rank of the object you pick is

$$s \frac{1 - e^{\lambda(s-1)}}{\lambda(1 - s)}.$$

What is the mean rank of your object?

- (b) What if X has the *logarithmic distribution*, $f_X(k) = cp^k/(k + 1); k \geq 1$?
- 15** A biased coin is tossed N times, where N is a Poisson random variable with parameter λ . Show that if H is the number of heads shown and T the number of tails, then H and T are independent Poisson random variables. Find the mean and variance of $H - T$.
- 16** A biased coin is tossed N times, where N is a random variable with finite mean. Show that if the numbers of heads and tails are independent, then N is Poisson. [You may want to use the fact that all continuous solutions of $f(x + y) = f(x)f(y)$ take the form $f(x) = e^{\lambda x}$ for some λ .]
- 17** Let X_n have a negative binomial distribution with parameters n and $p (= 1 - q)$. Show (using generating functions) that if $n \rightarrow \infty$ in such a way that $\lambda = nq$ remains constant, then $\lim_{n \rightarrow \infty} \mathbf{P}(X_n = k) = e^{-\lambda} \lambda^k / k!$. Show that $\mathbf{E}(X_n) = nqp^{-1}$ and $\text{var}(X_n) = nqp^{-2}$.
- 18** The events $(A_n; n \geq 1)$ are independent and $\mathbf{P}(A_n) = 1 - \exp(-\lambda_n)$. Define $N = \min\{n : A_n \text{ occurs}\}$.
- (a) Show that $\mathbf{E}(s^N) = s + (s - 1) \sum_{n=1}^{\infty} \exp\left(-\sum_{k=1}^n \lambda_k\right) s^n$.
- (b) Find $\mathbf{E}(s^N)$ and $\mathbf{E}(N)$, when $\lambda_n = a + \log n$.
- 19** The probability of obtaining heads when a certain coin is tossed is p . The coin is tossed repeatedly until a sequence of three heads is obtained. If p_n is the probability that this event occurs in n throws,

show that $p_0 = p_1 = p_2 = 0$, $p_3 = p^3$, and

$$p_n = p^3(1-p) \left(1 - \sum_{k=0}^{n-4} p_k \right) \quad \text{if } n \geq 4.$$

Show that the generating function $G(s) = \sum_{k=0}^{\infty} p_k s^k$ is given by

$$G(s) = \frac{p^3 s^3 (1 - ps)}{1 - s + p^3(1-p)s^4}.$$

Now find the expected number of throws of an unbiased coin needed to obtain three consecutive heads.

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

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

- 21** A gambler repeatedly plays the game of guessing whether a fair coin will fall heads or tails when tossed. For each correct prediction he wins £1, and for each wrong one he loses £1. At the start of play, he holds £ n (where n is a positive integer), and he has decided to stop play as soon as either (i) he has lost all his money, or (ii) he possesses £ K , where K is a given integer greater than n . Let $p(n)$ denote for $1 \leq n \leq K-1$ the probability that he loses all his money, and let $p(0) = 1$, $p(K) = 0$. Show that $p(n) = \frac{1}{2}(p(n-1) + p(n+1))$; ($1 \leq n \leq K-1$).

$$G(s) = \sum_{n=0}^{k-1} p(n)s^n$$

then, provided $s \neq 1$,

$$G(s) = \frac{1}{(1-s)^2} (1 - (2 - p(1))s + p(K-1)s^{K+1}).$$

Hence, or otherwise, show that $p(1) = 1 - 1/K$, $p(K-1) = 1/K$ and that, in general, $p(n) = 1 - n/K$.

- 22** A class of particles behaves in the following way. Any particle in existence at time n is replaced at time $n+1$ by a random number of similar particles having probability mass function $f(k) = 2^{-(k+1)}$, $k \geq 0$, independently of all other particles. At time zero, there is exactly one particle in existence and the set of all succeeding particles is called its descendants.

Let the total number of particles that have ever existed by time n be S_n . Show that the p.g.f. $G_n(z) = \mathbf{E}(z^{S_n})$ satisfies

$$G_n(z) = \frac{z}{2 - G_{n-1}(z)} \quad \text{for } 0 \leq z \leq 1 \text{ and } n \geq 1.$$

Deduce that with probability one, the number of particles that ever exist is finite, but that as $n \rightarrow \infty$, $\mathbf{E}(S_n) \rightarrow \infty$.

- 23** Let $G(s)$ be the generating function of the family size in an ordinary branching process. Let Z_n be the size of the population in the n th generation, and let T_n be the total number of individuals

who have ever lived up to that time. Show that $H_n(s, t)$, the joint generating function of Z_n and T_n satisfies $H_n(s, t) = tG(H_{n-1}(s, t))$.

- 24 Show that for each integer n , $(s + n - 1)(s + n - 2) \dots s/n!$, is the probability generating function of some random variable X . Show that as $n \rightarrow \infty$, $\mathbf{E}(X)/\log n \rightarrow 1$.
- 25 Find the probability generating function of the distribution

$$\mathbf{P}(X = k) = \left(\frac{a}{1+a}\right)^\lambda \frac{\lambda(\lambda+1)\dots(\lambda+k-1)}{(1+a)^k k!}; \quad k > 0, \mathbf{P}(X = 0) = \left(\frac{a}{1+a}\right)^\lambda.$$

- 26 Let X and Y be independent Poisson random variables with parameters λ and μ respectively. Find the joint probability generating function of $X - Y$ and $X + Y$.
Find the factorial moments of $X + Y$ and the cumulants of $X - Y$.
- 27 Let X be a binomial random variable with parameters n and p , and let Y be a binomial random variable with parameters m and $q (= 1 - p)$. Find the distribution of $X - Y + m$, and explain why it takes the form it does.
- 28 A series of objects passes a checkpoint. Each object has (independently) probability p of being defective, and probability α of being subjected to a check which infallibly detects a defect if it is present. Let N be the number of objects passing the checkpoint before the first defective is detected, and let D be the number of these passed objects that were defective (but undetected). Find:
(a) The joint p.g.f. of D and N .
(b) $\mathbf{E}(D/N)$.
- If the check is not infallible, but errs with probability δ , find the above two quantities in this case.
- 29 Let the sequence $(a_i; i \geq 0)$ be defined by

$$\frac{1}{(2-s)^{n+1}} = \sum_0^\infty a_i s^i.$$

Show that $\sum_{i=0}^n a_i = \frac{1}{2}$, and interpret this result in terms of random variables.

[Hint: $(1+x)^{n-r} = (1-x/(1+x))^r (1+x)^n$.]

- 30 A two-dimensional random walk $(X_n, Y_n; n \geq 0)$ evolves in the following way. If $(X_n, Y_n) = (x, y)$, then the next step is to one of the four points $(x+1, y)$, $(x-1, y)$, $(x, y+1)$, $(x, y-1)$, with respective probabilities $\alpha_1, \beta_1, \alpha_2, \beta_2$, where $\alpha_1 + \beta_1 + \alpha_2 + \beta_2 = 1$. Initially, $(X_0, Y_0) = (0, 0)$. Define $T = \min\{n; X_n + Y_n = m\}$.
Find the probability generating function of T .
- 31 In Problem 30, if $\alpha_1 = \beta_1$ and $\alpha_2 = \beta_2$, show that $\mathbf{E}(X_n^2 + Y_n^2) = n$.
Also, in Problem 30, if $\alpha_1 = \beta_1 = \alpha_2 = \beta_2 = \frac{1}{4}$,
(a) Show that $\mathbf{E}(T) = \infty$.
(b) Show that the point at which the walk hits $x + y = m$ is a proper random variable.
(c) Find its generating function $\mathbf{E}(s^{X_T - Y_T})$.
- 32 Use the identity $t(1+t)^{n-1} = (1+t)^n \sum_{i=0}^\infty (-t^{-1})^i$ to prove that

$$\binom{n}{i} - \binom{n}{i+1} + \dots + (-)^{n-i} \binom{n}{n} = \binom{n-1}{i-1}.$$

- 33 Let the generating function of the family size in an ordinary branching process be $G(s) = 1 - p(1-s)^\beta$; $0 < p, \beta < 1$. Show that if $Z_0 = 1$

$$\mathbf{E}(s^{Z_n}) = 1 - p^{1+\beta+\dots+\beta^{n-1}} (1-s)^{\beta^n}.$$

- 34 Let the generating function of the family size of an ordinary branching process be $G(s) = q + ps$, and let $\mathbf{E}(s^{Z_0}) = e^{\lambda(s-1)}$. Let $T = \min\{n; Z_n = 0\}$. Show that

$$\mathbf{P}(T = n) = e^{-\lambda p^{n+1}} - e^{-\lambda p^n}.$$

- 35 Let the number of tosses required for a fair coin to show a head be T . An integer X is picked at random from $\{1, \dots, T\}$ with equal probability $\frac{1}{T}$ of picking any one. Find $G_X(s)$.
- 36 Show that for $\alpha > 0, \beta > 0, \alpha + \beta < 1$,

$$G(s, t) = \frac{\log(1 - \alpha s - \beta t)}{\log(1 - \alpha - \beta)}$$

is a bivariate p.g.f. Find the marginal p.g.f.s and the covariance.

- 37 Let X and Y be independent with r th cumulants $\kappa_r^{(X)}$ and $\kappa_r^{(Y)}$. Show that $X + Y$ has r th cumulant $\kappa_r^{(X)} + \kappa_r^{(Y)}$.
- 38 Let X have cumulants $\kappa_r; r \geq 1$ and moments $\mu_r; r \geq 1$. Show that $\kappa_1 = \mathbf{E}(X), \kappa_2 = \text{var}(X)$, and $\kappa_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3$.
- 39 Show that the joint probability mass function

$$f(x, y) = \frac{1 - \lambda - \mu}{\mu} \binom{x + y - 1}{x} \lambda^x \mu^y; \quad x \geq 0, y \geq 1$$

has joint p.g.f.

$$G(s, t) = \frac{(1 - \lambda - \mu)t}{1 - \lambda s - \mu t}.$$

What is $\text{cov}(X, Y)$?

- 40 Let X_m have generating function $(p/(1 - qs))^m$, where $p = 1 - q > 0$. Show that as $m \rightarrow \infty$

$$\mathbf{E}(s^{X_m} | X_m > 0) \rightarrow \frac{\log(1 - qs)}{\log(1 - q)}.$$

- 41 Prove the identity

$$\sum_{k=0}^n \binom{2k}{k} 4^{-k} = (2n + 1) \binom{2n}{n} 4^{-n}.$$

Now let S_n be a simple symmetric random walk with $S_0 = 0$. Let v_n be the expected number of visits of the walk to zero, up to and including time n . Show that (including the initial visit)

$$v_{2n} = v_{2n+1} = (2n + 1) \binom{2n}{n} 2^{-2n}.$$

- 42 Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = 0$. Let R_r be the number of steps until the walk first revisits the origin for the r th time, and let T_{0r} be the number of steps until the walk first visits r . Show that

$$\mathbf{E}(s^{T_{0r}}) = \left(\frac{1}{2qs}\right)^r \mathbf{E}(s^{R_r}).$$

Deduce that $\mathbf{P}(R_r = n) = (2q)^r \mathbf{P}(T_r = n - r)$, and hence that, as $n \rightarrow \infty$

$$\frac{\mathbf{P}(R_r = n)}{\mathbf{P}(R_1 = n)} \rightarrow r.$$

[H. Kesten and F. Spitzer have shown that this remains true for a wider class of random walks. (*J. d'Anal. Math.* Vol. 11, 1963).]

- 43** Let X be geometric with parameter p . Use the fact that $q\mathbf{E}s^X = \mathbf{E}s^{X-1} - p$ to deduce that $\mathbf{E}(X - 1)^k = q\mathbf{E}X^k$ for $k \geq 1$.
- 44** If X has p.g.f. $G(s)$, show that $T(s) = \sum_s^n s^n \mathbf{P}(X > n) = (1 - G(s))/(1 - s)$. Deduce that $\mathbf{E}X(X - 1) \dots (X - k + 1) = kT^{(k-1)}(1)$.

7

Continuous Random Variables

He talks at random: sure the man is mad.
W. Shakespeare, Henry VI

7.1 Density and Distribution

Hitherto, we have assumed that a random variable can take any one of only a countable set of values. However, suppose your height is 5 feet or 6 feet (or somewhere in between). Then, previously (however briefly), your height in feet has taken every value in $[1, 5]$, including $\sqrt{2}$, e , π , and so on. (Each value can be taken more than once because you are taller in the morning than in the evening.) Thus, if X is the height of a randomly selected member of the population, the state space of X is not countable. There are many other simple examples of variables that may take any one of an uncountable number of values; for example, the brightness of a randomly chosen star, the time until some cell divides, the velocity of a comet, the direction of the wind, and so on. Think of some yourself.

In view of these remarks, we are about to introduce a new class of random variables such that the state space Ω is uncountable, and $X(\omega)$ may take any one of an uncountable number of real values. However, before we embark on this task, it is as well to reassure you that, despite their separate presentation, these new random variables share most of the useful properties of discrete random variables. Also, many of these properties are proved in exactly the same way as in the discrete case and (even better) we are able to use much of the same notation.

Thus, as in the discrete case, we start with a probability function $P(\cdot)$ defined on a collection \mathcal{F} (the event space) of subsets of Ω (the sample space). Then we think of a random variable X as a real valued function $X(\omega)$ defined for each $\omega \in \Omega$.

Our first requirement (as in the discrete case) is a function that tells us about the relative likelihoods of possible values of X . Happily, we already have such a function; recall the following:

(1) **Definition** The distribution function F of the random variable X is the function

(2)
$$F(x) = \mathbf{P}(A_x),$$

where A_x is the event $A_x = \{\omega : X(\omega) \leq x\}$, $x \in \mathbb{R}$. ▲

We usually write (2) as

(3)
$$F(x) = \mathbf{P}(X \leq x),$$

and denote F by $F_X(x)$ when we want to stress the role of X .

Notice that for Definition 1 to be meaningful, the event A_x must be in \mathcal{F} , so that we know $\mathbf{P}(A_x)$. Thus, $X(\cdot)$ only qualifies to appear in Definition 1 if $A_x \in \mathcal{F}$ for all x . This is true throughout this book; the implications of this so-called *measurability* condition are explored in more advanced books.

(4) **Example: Uniform Distribution** You devise an experiment in which the outcome is equally likely to be any point Q in the interval $[0, 1]$. Thus, the sample space is the set of points $(Q : Q \in [0, 1])$. The event space \mathcal{F} will include all intervals in $[0, 1]$; we omit the proof that such an \mathcal{F} exists, but you should rest assured that it does exist. Define the random variable $X(Q)$ to be the distance from the origin O to Q . From the nature of the experiment, if

$$A_{ab} = \{Q : Q \in (a, b)\}, \quad 0 \leq a \leq b \leq 1,$$

then $\mathbf{P}(A_{ab}) = b - a$. Hence, X has distribution function sketched in Figure 7.1

$$F_X(x) = \mathbf{P}(A_{0x}) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1. \end{cases} \quad \bullet$$

In the future, the underlying sample space will make few appearances. We tend to think of the possible values of X as the sample space, as we did for discrete random variables. (It can be proved that this is a permissible view.)

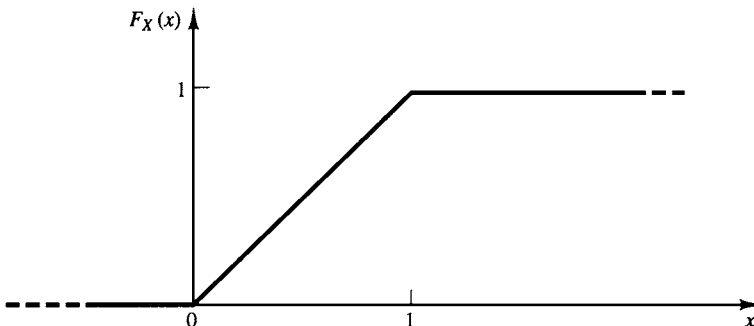


Figure 7.1 The distribution function $F_X(x)$ of a random variable X distributed uniformly on $(0, 1)$.

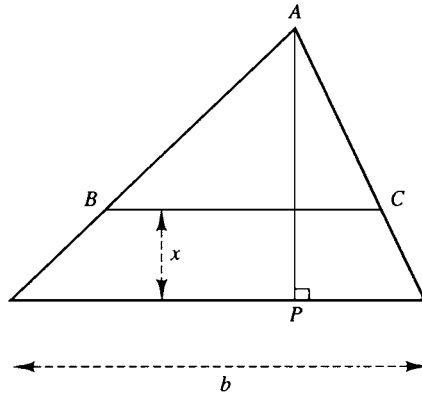


Figure 7.2 The altitude AP is of length h .

- (5) **Example** A point Q is picked at random in a triangle of area a , with base of length b . Let X be the perpendicular distance from Q to the base. What is the distribution function $F_X(x)$? [See figure 7.2 for a sketch of the triangle].

Solution Let the height of the triangle AP be h . The event $X > x$ occurs when Q lies inside the triangle ABC . For reasons of symmetry,

$$\mathbf{P}(Q \in ABC) = (\text{area of } ABC)/a = \left(\frac{h-x}{h}\right)^2.$$

Hence,

$$F_X(x) = \mathbf{P}(X \leq x) = 1 - \mathbf{P}(Q \in ABC) = \begin{cases} 0 & x < 0 \\ 1 - \left(\frac{h-x}{h}\right)^2 & 0 \leq x \leq h \\ 1 & x > h. \end{cases} \bullet$$

We summarize the basic properties of $F(x)$ in the following:

- (6) **Theorem** Let X have distribution function $F(x)$. Then $0 \leq F(x) \leq 1$ for all x , and

(7)
$$\mathbf{P}(x < X \leq y) = F(y) - F(x) \geq 0, \text{ for all } x \leq y.$$

- (8) If $h > 0$, then

$$\lim_{h \rightarrow 0} F(x+h) = F(x).$$

- (9) If $\mathbf{P}(|X| < \infty) = 1$, then

$$\lim_{x \rightarrow \infty} F(x) = 1 \quad \text{and} \quad \lim_{x \rightarrow -\infty} F(x) = 0.$$

In plain words, this theorem says that as x increases, $F(x)$ is nondecreasing, continuous on the right, and lies between 0 and 1. It can be shown conversely that any function with these properties is the distribution function of some random variable.

Proof The first result follows from (2) because $0 \leq \mathbf{P}(A_x) \leq 1$. To show (7), note that

$$\{\omega: x < X \leq y\} = A_y \cap A_x^c = A_y \setminus A_x.$$

Because $A_x \subseteq A_y$ for $x \leq y$, we have

$$\mathbf{P}(x < X \leq y) = \mathbf{P}(A_y) - \mathbf{P}(A_x) = F(y) - F(x) \geq 0$$

by the nonnegativity of $\mathbf{P}(\cdot)$. To prove (8), we use the continuity of $\mathbf{P}(\cdot)$, see (1.5.4.). Let $(h_k; k \geq 1)$ be any sequence decreasing to zero, and let $A_{(k)}$ be the event that $X \leq x + h_k$. Then

$$\begin{aligned} \lim_{h \rightarrow 0} F(x + h) &= \lim_{k \rightarrow \infty} F(x + h_k) = \lim_{k \rightarrow \infty} \mathbf{P}(A_{(k)}) \\ &= \mathbf{P}(\lim_{k \rightarrow \infty} A_{(k)}) = \mathbf{P}(A_x) = F(x), \end{aligned}$$

as required. Finally, for (9),

$$\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} \mathbf{P}(A_n) = \mathbf{P}(\Omega) \quad \text{by (1.5.4).}$$

The last part is proved similarly. ■

Although the distribution function has not played a very active role so far in this book, it now assumes a greater importance. One reason for this is Theorem 6 above, which shows that $F(x)$ really is a function that can tell us how likely X is to be in some simple subset of the real line. Another reason is the following simple corollary of Example 5.

(10) Corollary *If $F(x)$ is continuous, then for all x $\mathbf{P}(X = x) = 0$.*

Proof

$$\begin{aligned} \mathbf{P}(X = x) &= \lim_{n \rightarrow \infty} \mathbf{P}\left(x - \frac{1}{n} < X \leq x\right) \quad \text{by (1.5.4)} \\ &= \lim_{n \rightarrow \infty} \left(F(x) - F\left(x - \frac{1}{n}\right)\right) \quad \text{by Theorem 6} \\ &= 0 \quad \text{because } F \text{ is continuous.} \end{aligned} \quad \blacksquare$$

(11) Example Let X be uniformly distributed on $(0, 1)$. Then

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

is clearly continuous, so that $\mathbf{P}(X = x) = 0$ for all x . ■

If F_X is continuous, then X is known as a continuous random variable. We now define a particularly important class of continuous random variables.

(12) **Definition** Let X have distribution function F . If the derivative

$$(13) \quad \frac{dF}{dx} = F'(x)$$

exists at all but a finite number of points, and the function f defined by

$$(14) \quad f(x) = \begin{cases} F'(x), & \text{where } F'(x) \text{ exists} \\ 0 & \text{elsewhere,} \end{cases}$$

satisfies

$$(15) \quad F(x) = \int_{-\infty}^x f(v) dv,$$

then X is said to be a continuous random variable with density $f(x)$. ▲

It follows from (15) that if X has density $f(x)$, then for $C \subseteq \mathbb{R}$ the Key Rule is:

$$\mathbf{P}(X \in C) = \int_C f(x) dx,$$

when both sides exist. [In line with the remark following (3), they exist if $\{\omega: X(\omega) \in C\} \in \mathcal{F}$.]

Example 4 Revisited: Uniform Density

uniformly at random in $[0, 1]$, then

In Example 4, we found that if X was chosen

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1. \end{cases}$$

It follows that X has a density

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases} \bullet$$

Figures 7.3 to 7.6 illustrate the density and distribution of random variables Y and Z , which are uniformly distributed on (a, b) and $(a, b) \cup (c, d)$, respectively.

Example (5) Revisited

For the point Q picked at random in a triangle, where X is the perpendicular distance from Q to the base, we found

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - \left(\frac{h-x}{h}\right)^2 & 0 \leq x \leq h \\ 1 & x > h. \end{cases}$$

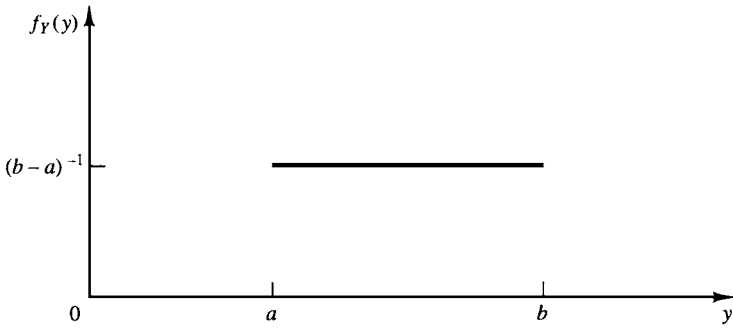


Figure 7.3 The density function $f(y)$ of a random variable distributed uniformly on (a, b) .

It follows that X has density

$$f_X(x) = \begin{cases} 2(h-x)/h^2 & 0 < x < h \\ 0 & \text{otherwise.} \end{cases}$$

There are continuous random variables that do not have a density, but none appear in this text. Therefore, whenever X is said to be continuous here, it always follows that X has a density f .

Let us consider a basic example. In the discrete case, the geometric distribution is of the form $F(x) = 1 - q^x$, if $x = 1, 2, \dots$. It is natural to consider the analogous distribution in the continuous case, which turns out to be equally (if not more) important.

(16) Example: Exponential Density Let X have distribution

$$F(x) = \max\{0, 1 - e^{-\lambda x}\},$$

where $\lambda > 0$. Then F is continuous for all x , and also differentiable except at $x = 0$, so

$$F'(x) = \begin{cases} 0 & \text{if } x < 0, \\ \lambda e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

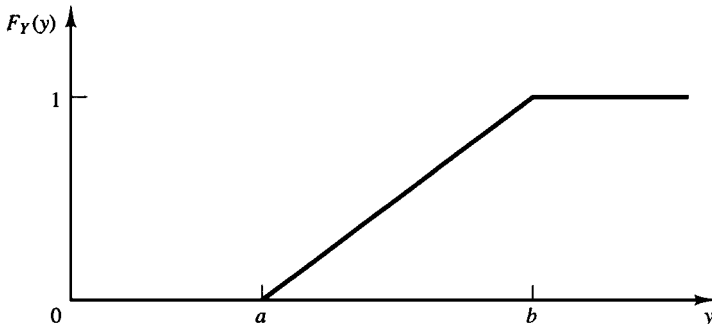


Figure 7.4 The distribution function $F(y)$ of a random variable distributed uniformly on (a, b) .

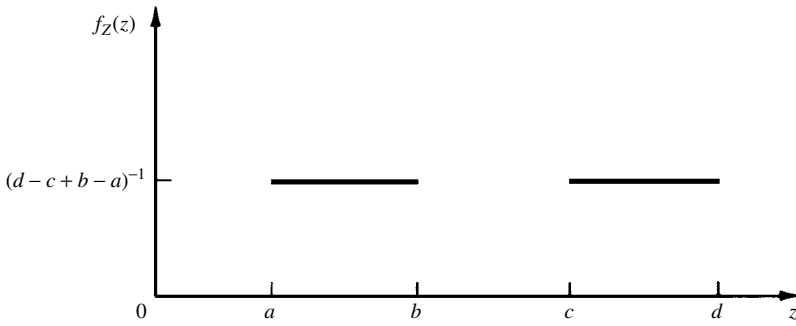


Figure 7.5 The density function of a random variable distributed uniformly on $(a, b) \cup (c, d)$, where $a < b < c < d$.

Now let

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases}$$

Then of course, for $x > 0$,

$$F(x) = 1 - e^{-\lambda x} = \int_0^x \lambda e^{-\lambda v} dv = \int_{-\infty}^x f(v) dv,$$

and, for $x < 0$,

$$F(x) = 0 = \int_{-\infty}^x f(v) dv.$$

Hence, $f(x)$ is a density of X . See Figures 7.7 and 7.8. ●

Notice that $F'(0)$ does not exist, and also that the function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

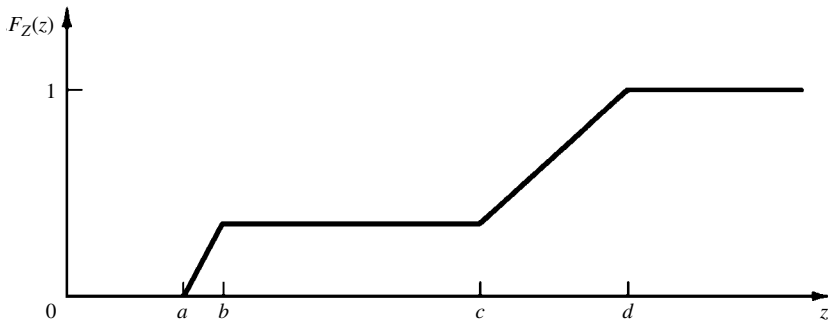


Figure 7.6 The distribution function of a random variable distributed uniformly on $(a, b) \cup (c, d)$, where $a < b < c < d$.

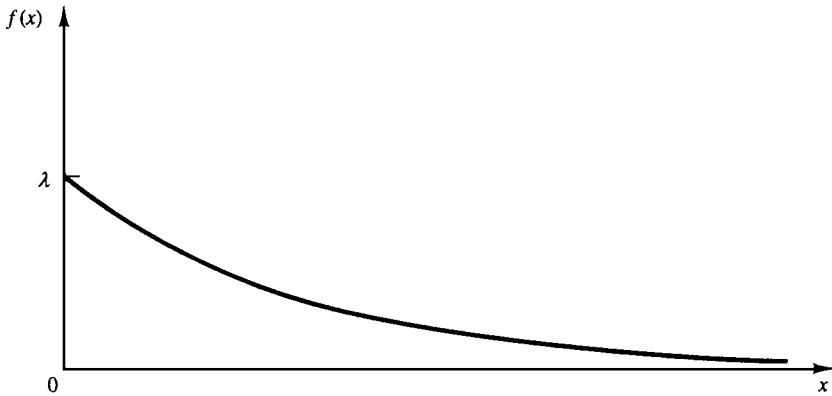


Figure 7.7 The density of an exponential random variable with parameter λ ; $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$.

would also satisfy $F(x) = \int_{-\infty}^x f(v)dv$. This illustrates the fact that (15) does not uniquely determine $f(x)$ given $F(x)$. However, this problem can only arise at a finite number of points and (to some extent) it does not matter what value we give $f(x)$ at any of these exceptional points. Usually we make it zero, but consider the following:

(17) Example: Two-Sided Exponential Distribution Let X have distribution

$$F(x) = \begin{cases} pe^{\lambda x} & \text{if } x < 0 \\ 1 - (1 - p)e^{-\lambda x} & \text{if } x \geq 0, \end{cases}$$

where $0 < p < 1$ and $\lambda > 0$. Then $F(x)$ is continuous and

$$F'(x) = \begin{cases} \lambda pe^{\lambda x} & \text{if } x < 0 \\ \lambda(1 - p)e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

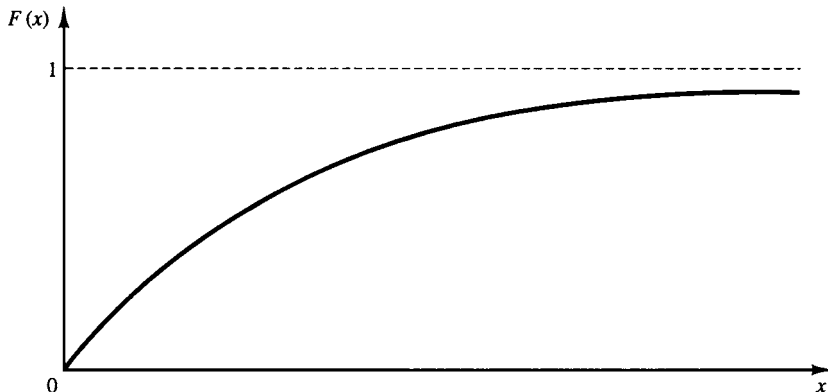


Figure 7.8 The distribution of an exponential random variable with parameter λ ; $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$.

A suitable density is

$$f(x) = \begin{cases} \lambda p e^{\lambda x} & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ \lambda(1-p)e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

However, if $p = \frac{1}{2}$, it is tempting to set $f(0) = \frac{\lambda}{2}$ and write $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$ for all x . The point is that it really does not matter very much. ●

Finally, we note the obvious facts that for any density f ,

$$(18) \quad f(x) \geq 0,$$

and if $\mathbf{P}(|X| < \infty) = 1$, then

$$(19) \quad \int_{-\infty}^{\infty} f(v)dv = 1.$$

It is straightforward to see that any function with these properties is the density of some random variable, and so any integrable nonnegative function can be used to form a density.

Example The function $g(x) = x^2 - x + 1$ is easily seen to be nonnegative for all x , and

$$\int_{+a}^{+b} g(v)dv = \frac{1}{3}(b^3 - a^3) - \frac{1}{2}(b^2 - a^2) + b - a = c(a, b), \quad \text{say.}$$

Hence, the function

$$f(x) = \begin{cases} c(a, b)^{-1}g(x) & \text{if } a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

is a density function. ●

(20) **Example: Cauchy Distribution** Show that for an appropriate choice of the constant $c(a, b)$, the function

$$f(x) = \begin{cases} c(a, b)(1+x^2)^{-1} & \text{if } a < x < b \\ 0 & \text{elsewhere} \end{cases}$$

is a density function. Show that $c(-\infty, \infty) = 1/\pi$ and that $c(-1, 1) = c(0, \infty) = 2/\pi$.

Solution Trivially, $f \geq 0$ if $c \geq 0$. New recall that

$$\frac{d}{dx} \tan^{-1} x = (1+x^2)^{-1}.$$

Thus, $\int_a^b f(x)dx = 1$ if and only if $c(a, b)^{-1} = \tan^{-1} b - \tan^{-1} a$. In particular,

$$c(-\infty, \infty)^{-1} = \frac{\pi}{2} + \frac{\pi}{2} = \pi,$$

as required, and

$$c(-1, 1)^{-1} = \frac{\pi}{4} + \frac{\pi}{4} = c(0, \infty)^{-1}.$$

In general then, given a nonnegative function $g(x)$, the function

$$f(x) = g(x) \left(\int_{-\infty}^{\infty} g(v) dv \right)^{-1}$$

is a density, if the integral exists. To discover whether it does, the following technique is useful. If we can find a constant b such that for all $n > 0$, $\int_0^n g(v) dv < b < \infty$, then

$$\int_0^{\infty} g(v) dv = \lim_{n \rightarrow \infty} \int_0^n g(v) dv$$

exists by monotone convergence.

(21) **Example: Normal Density** Show that

$$f = c \exp\left(-\frac{1}{2}x^2\right) \quad \text{for all } x \in \mathbb{R}$$

can be a density.

Solution For any $n > 1$,

$$\int_{-n}^n \exp\left(-\frac{1}{2}v^2\right) dv < 2 \left(\int_0^1 dv + \int_1^n e^{-v} dv \right) < 2(1 + e^{-1}).$$

Hence, $c^{-1} = \int_{-\infty}^{\infty} \exp(-\frac{1}{2}v^2) dv$ exists, and f is a density for this c .

Remark In fact, it can be shown that

$$(22) \quad c^{-1} = (2\pi)^{1/2}.$$

The proof of this is not quite trivial; we give it in Example 8.3.8. Also, note that there are other normal densities; the one in Example 21 is called the *standard* normal density denoted by $N(0, 1)$, and by

$$(23) \quad \phi(x) = (2\pi)^{-1/2} \exp\left(-\frac{1}{2}x^2\right).$$

Its distribution is $\Phi(x)$, given by

$$(24) \quad \Phi(x) = \int_{-\infty}^x \phi(v) dv.$$

(25) **Example: Gamma Distribution** Show that for $\alpha, \lambda, x > 0$, the function $f(x) = c\lambda^\alpha x^{\alpha-1} e^{-\lambda x}$ can be a density. When α is a positive integer, show that $c^{-1} = (\alpha - 1)!$

Solution First, we show that the integral $\int_0^\infty x^{\alpha-1} e^{-\lambda x} dx$ exists. Because $x^{\alpha-1} e^{-\frac{1}{2}\lambda x} \rightarrow 0$ as $x \rightarrow \infty$, there is some $m < \infty$ such that for $x > m$, $x^{\alpha-1} e^{-\frac{1}{2}\lambda x} < 1$. Hence, for $n > m$,

$$\begin{aligned} \int_0^n v^{\alpha-1} e^{-\lambda v} dv &< \int_0^m v^{\alpha-1} e^{-\lambda v} dv + \int_m^n e^{-\frac{1}{2}\lambda v} dv \\ &< \int_0^m v^{\alpha-1} e^{-\lambda v} dv + 2\lambda^{-1} e^{-\frac{1}{2}\lambda m} = b \text{ (say)}. \end{aligned}$$

Hence, $c^{-1} = \int_0^\infty \lambda^\alpha x^{\alpha-1} e^{-\lambda x} dx$ exists, and if α is a positive integer then repeated integration by parts gives $c^{-1} = (\alpha - 1)! \int_0^\infty \lambda e^{-\lambda v} dv = (\alpha - 1)!$ ●

This density is known as the *gamma density* with parameters α and λ . When α is not an integer, the integral above defines a function of α known as the *gamma function*, and denoted by

$$(26) \quad \Gamma(\alpha) = \int_0^\infty \lambda^\alpha v^{\alpha-1} e^{-\lambda v} dv = \int_0^\infty u^{\alpha-1} e^{-u} du.$$

The density function in this case is

$$(27) \quad f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} \text{ for } x > 0.$$

In particular, when $\lambda = \frac{1}{2}$ and $\alpha = \frac{n}{2}$, the density

$$(28) \quad f(x) = \frac{1}{\Gamma(\frac{n}{2})} \lambda^{\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{1}{2}x}$$

is known as the $\chi^2(n)$ density, and it is referred to as *chi-squared* with n degrees of freedom.

We conclude with another way of producing densities.

(29) Example: Mixtures Let $f_1(x)$ and $f_2(x)$ be density functions, and let $f_3(x) = \gamma f_1(x) + (1 - \gamma) f_2(x)$, where $0 \leq \gamma \leq 1$. Then $f_3(x) \geq 0$, and

$$\int f_3 = \gamma \int f_1 + (1 - \gamma) \int f_2 = 1.$$

Hence, f_3 is a density, and is said to be a *mixture* of f_1 and f_2 . ●

For example, the two-sided exponential density of Example 17 may now be seen as a mixture of $f_1 = \lambda e^{\lambda x} (x < 0)$ and $f_2 = \lambda e^{-\lambda x} (x > 0)$ with $\gamma = p$.

7.2 Functions of Random Variables

Suppose that X and Y are random variables such that $Y = g(X)$, where $g(\cdot)$ is some given function. If we know the density of X , can we find the distribution of Y ? In general terms, the answer is straightforward because, by the properties of densities and

distributions,

$$(1) \quad F(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \int_C f_X(v) dv,$$

where $C = \{v : g(v) \leq y\}$. Then, if $F(y)$ is continuous and differentiable, we can go on to find the density of Y , if it exists. Here are some simple examples of this idea in practice.

(2) **Example** Let X be uniformly distributed on $(0, 1)$ with density

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $Y = -\lambda^{-1} \log X$, where $\lambda > 0$, what is the density of Y ?

Solution First, we seek the distribution of Y :

$$\begin{aligned} F_Y(y) &= \mathbf{P}(-\lambda^{-1} \log X \leq y) = \mathbf{P}(\log X \geq -\lambda y) \\ &= \mathbf{P}(X \geq \exp(-\lambda y)) = \begin{cases} 1 - e^{-\lambda y} & \text{for } y \leq 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Hence, the derivative exists except at $y = 0$, and

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y > 0 \\ 0 & \text{if } y \leq 0. \end{cases}$$

This is the exponential density with parameter λ . ●

Some care is required if $g(\cdot)$ is not one–one.

(3) **Example** Let X be uniformly distributed on $[-1, 1]$. Find the density of $Y = X^r$ for nonnegative integers r .

Solution First, note that X has distribution function $F(x) = \frac{1}{2}(1 + x)$ for $-1 \leq x \leq 1$. Now, if r is odd, then the function $g(x) = x^r$ maps the interval $[-1, 1]$ onto itself in one–one correspondence. Hence, routinely:

$$\mathbf{P}(Y \leq y) = \mathbf{P}(X^r \leq y) = \mathbf{P}(X \leq y^{1/r}) = \frac{1}{2}(1 + y^{1/r}) \quad \text{for } -1 \leq y \leq 1,$$

and Y has density

$$f(y) = \frac{1}{2r} y^{\frac{1}{r}-1}, \quad -1 \leq y \leq 1.$$

If r is even, then $g(x) = x^r$ takes values in $[0, 1]$ for $x \in [-1, 1]$. Therefore,

$$\mathbf{P}(Y \leq y) = \mathbf{P}(0 \leq X^r \leq y) = \mathbf{P}(-y^{1/r} \leq X \leq y^{1/r}) = y^{1/r} \quad \text{for } 0 \leq y \leq 1.$$

Hence, Y has density

$$f(y) = \frac{1}{r} y^{\frac{1}{r}-1}, \quad 0 \leq y \leq 1.$$

Finally, if $r = 0$, then $X^r = 1$, $F_Y(y)$ is not continuous (having a jump from 0 to 1 at $y = 1$) and so Y does not have a density in this case. Obviously, Y is discrete, with $\mathbf{P}(Y = 1) = 1$. ●

(4) **Example** Let X have the standard normal distribution with density

$$f(x) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^2\right).$$

Find the density of $Y = \sigma X + \mu$ for given constants μ and $\sigma \neq 0$. Also, find the density of $Z = X^2$.

Solution Adopting the by now familiar technique:

$$(5) \quad \mathbf{P}(\sigma X + \mu \leq y) = \mathbf{P}(\sigma X \leq y - \mu) = \begin{cases} \mathbf{P}\left(X \leq \frac{y - \mu}{\sigma}\right) & \text{if } \sigma > 0 \\ \mathbf{P}\left(X \geq \frac{y - \mu}{\sigma}\right) & \text{if } \sigma < 0 \end{cases}$$

$$= \begin{cases} F_X\left(\frac{y - \mu}{\sigma}\right) & \text{if } \sigma > 0 \\ 1 - F_X\left(\frac{y - \mu}{\sigma}\right) & \text{if } \sigma < 0 \end{cases}$$

Hence, differentiating (5) with respect to y ,

$$(6) \quad f_Y(y) = \frac{1}{|\sigma|} f_X\left(\frac{y - \mu}{\sigma}\right) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\frac{y - \mu}{\sigma}\right)^2\right).$$

Second,

$$\mathbf{P}(X^2 \leq z) = \mathbf{P}(X \leq \sqrt{z}) - \mathbf{P}(X \leq -\sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z}).$$

Differentiating now gives

$$(7) \quad f_Z(z) = \frac{1}{2\sqrt{z}} f_X(\sqrt{z}) + \frac{1}{2\sqrt{z}} f_X(-\sqrt{z}) = \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{1}{2}z\right). \quad \bullet$$

Remark The density given by (6) is known as the *normal* density with parameters μ and σ^2 , sometimes denoted by $N(\mu, \sigma^2)$. The standard normal density of Example 7.1.21 was $N(0, 1)$ because $\phi(x)$ has $\mu = 0$ and $\sigma = 1$.

The density given by (7) is the gamma density of (7.1.23) with parameters $\frac{1}{2}$ and $\frac{1}{2}$. This is known as the chi-squared density with parameter 1, sometimes denoted by $\chi^2(1)$. This is a special case of (7.1.28).

(8) **Example: Inverse Functions** Let X have distribution function $F(x)$, where $F(x)$ is continuous and strictly increasing. Let $g(x)$ be a function satisfying $F(g(x)) = x$. Because

$F(x)$ is continuous and strictly increasing, this defines $g(x)$ uniquely for every x in $(0, 1)$. The function $g(\cdot)$ is called the inverse function of $F(\cdot)$ and is often denoted by

$$g(x) = F^{-1}(x).$$

Clearly, F is the inverse function of g , that is

$$(9) \quad g(F(x)) = F(g(x)) = x,$$

and $g(x)$ is an increasing function.

(a) Use this function to show that $Y = F(X)$ is uniformly distributed on $(0, 1)$.

(b) Show that if U is uniform on $(0, 1)$, then $Z = F^{-1}(U)$ has distribution $F(z)$.

Solution (a) As usual, we seek the distribution function

$$\begin{aligned} \mathbf{P}(Y \leq y) &= \mathbf{P}(F(X) \leq y) = \mathbf{P}(g(F(X)) \leq g(y)) = \mathbf{P}(X \leq g(y)) && \text{by (9)} \\ &= F(g(y)) = y && \text{by (9)}. \end{aligned}$$

(b) Again,

$$\begin{aligned} \mathbf{P}(F^{-1}(U) \leq z) &= \mathbf{P}(F(g(U)) \leq F(z)) = \mathbf{P}(U \leq F(z)) && \text{by (9)} \\ &= F(z). && \bullet \end{aligned}$$

Although we have introduced them separately, discrete and continuous variables do have close links. Here are some examples to show this.

(10) **Example: Step Functions** Let X have distribution function $F(x)$ and density f . Define the function $S : \mathbb{R} \rightarrow \mathbb{Z}$ by

$$(11) \quad S(X) = k, \text{ if } k \leq X < k + 1,$$

where k is any integer. Then $S(X)$ is an integer valued discrete random variable with mass function

$$(12) \quad f_S(k) = \mathbf{P}(k \leq X < k + 1) = \int_k^{k+1} f(v)dv.$$

Obviously, $\mathbf{P}(S(X) \leq X) = 1$ and $F_S(x) \geq F_X(x)$, and

$$(13) \quad |S(X) - X| \leq 1. \quad \bullet$$

Now equation (13) shows that the integer valued $S(X)$ is, in some sense, a rough approximation to the continuous random variable X . It is easy to get much better approximations as follows.

(14) **Example: Discrete Approximation** As usual, X has density $f(x)$; suppose also that $X > 0$. For fixed n , with $0 \leq r \leq 2^n - 1$ and $k \geq 0$, define

$$S_n(X) = k + r2^{-n} \text{ if } k + r2^{-n} \leq X < k + (r + 1)2^{-n}.$$

Then $S_n(X)$ is a discrete random variable taking values in $(k + r2^{-n}; k \geq 0, 0 \leq r \leq 2^n - 1)$, with

$$\mathbf{P}(S_n(X) = k + r2^{-n}) = \int_{k+r2^{-n}}^{k+(r+1)2^{-n}} f(v)dv.$$

Again, we have $S_n(X) \leq X$, but this time, by the construction,

$$(15) \quad |S_n(X) - X| \leq 2^{-n}.$$

Thus, by choosing n large enough, we can find a discrete random variable $S_n(X)$ such that $|X - S_n(X)|$ is as small as we please. ●

In fact, it can be shown that we can find a simple random variable (taking only a finite number of values) that is arbitrarily close to X , but in a weaker sense than (15). (See Problem 12.)

7.3 Simulation of Random Variables

A random variable is a mathematical concept (having no other existence) that is suggested by the outcomes of real experiments. Thus, tossing a coin leads us to define an $X(\cdot)$ such that $X(H) = 1$, $X(T) = 0$, and X is the number of heads. The coin exists, X is a concept. A natural next step, having developed theorems about mathematical coins (e.g., the arc-sine laws) is to test them against reality. However, the prospect of actually tossing a large enough number of coins to check the arc-sine laws is rather forbidding.

Luckily, we have machines to do large numbers of boring and trivial tasks quickly, namely, computers. These can be persuaded to produce many numbers $(u_i; i \geq 1)$ that are sprinkled evenly and “randomly” over the interval $(0, 1)$. The word randomly appears in quotations because each u_i is not really random. Because the machine was programmed to produce it, the outcome is known in advance, but such numbers behave for many practical purposes as though they were random. They are called pseudorandom numbers.

Now if we have a pseudorandom number u from a collection sprinkled uniformly in $(0, 1)$, we can look to see if $u < \frac{1}{2}$, in which case we call it “heads”, or $u > \frac{1}{2}$ in which case we call it “tails.” This process is called *simulation*; we have simulated tossing a coin.

Different problems produce different random variables, but computers find it easiest to produce uniform pseudorandom numbers. We are thus forced to consider appropriate transformations of uniform random variables, and therefore many of the results of Section 7.2 find concrete applications when we seek to simulate random variables. A natural first question (before “how”) is *why* might we want to simulate such random variables? Some examples should suffice to answer this question.

- (1) **Example: Epidemic** An infection is introduced into a population. For each individual the incubation period is a random variable X , the infectious period is a random variable Y , and the number of further individuals infected is a random variable N , depending on Y and the behaviour of the infected individual. What happens? Unfortunately, exact solutions

to such problems are rare and, for many diseases (e.g., the so-called “slow viruses” or prions), X and Y are measured in decades so experiments are impractical. However, if we could simulate X and Y and the infection process N , then we could produce one simulated realization (not a real realization) of the epidemic. With a fast computer, we could do this many times and gain a pretty accurate idea of how the epidemic would progress (if our assumptions were correct). ●

- (2) **Example: Toll Booths** Motorists are required to pay a fee before entering a toll road. How many toll booths should be provided to avoid substantial queues? Once again an experiment is impractical. However, simple apparatus can provide us with the rates and properties of traffic on equivalent roads. If we then simulate the workings of the booth and test it with the actual traffic flows, we should obtain reasonable estimates of the chances of congestion. ●

Because of the ready availability of large numbers of uniform pseudorandom numbers, interest is concentrated on finding transformations that then yield random variables of arbitrary type. We have seen several in Section 7.2. Here is another idea.

Example: Composition The pseudorandom variable U is uniformly distributed on $(0, 1)$. Show how to simulate a random variable with density

$$f_X = \frac{1}{4}(x^{-\frac{1}{2}} + (1-x)^{-\frac{1}{2}}), \quad 0 < x < 1.$$

Solution Recall that if U is uniform on $(0, 1)$ then U^2 has density $f_1(x) = \frac{1}{2}x^{-\frac{1}{2}}$. Next consider

$$\mathbf{P}(1 - U^2 \leq x) = \mathbf{P}(U \geq (1-x)^{\frac{1}{2}}) = 1 - (1-x)^{\frac{1}{2}}.$$

Hence, $1 - U^2$ has density $f_2(x) = \frac{1}{2}(1-x)^{-\frac{1}{2}}$. Now toss a coin (real or simulated), and write

$$X = \begin{cases} U^2 & \text{if it's heads} \\ 1 - U^2 & \text{if it's tails.} \end{cases}$$

Then

$$f_X(x) = \frac{1}{2}f_1(x) + \frac{1}{2}f_2(x) = \frac{1}{4}(x^{-\frac{1}{2}} + (1-x)^{-\frac{1}{2}}),$$

as required. ●

We describe other methods of simulation as the necessary ideas are developed.

7.4 Expectation

Random variables with a density may have an expected value, similar to random variables with a mass function.

(1) **Definition** Let X have density $f(x)$. If $\int_{-\infty}^{\infty} |v|f(v)dv < \infty$, then X has an expected value, which is given by

$$(2) \quad \mathbf{E}(X) = \int_{-\infty}^{\infty} vf(v)dv. \quad \blacktriangle$$

(3) **Example: Uniform Density** Let X be uniformly distributed on (a, b) . Then

$$\mathbf{E}(X) = \int_a^b \frac{v}{b-a} dv = \frac{1}{2}(b-a). \quad \bullet$$

(4) **Example: Exponential Density** Let X have density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Then

$$\mathbf{E}(X) = \int_0^{\infty} v\lambda e^{-\lambda v} dv = \lambda^{-1}. \quad \bullet$$

(5) **Example: Normal Density** Let X have the $N(\mu, \sigma^2)$ density. Then

$$\begin{aligned} \mathbf{E}(X) &= \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} v \exp(-(v-\mu)^2/(2\sigma^2)) dv \\ &= \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} (v-\mu) \exp\left(-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right) dv \\ &\quad + \frac{\mu}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right) dv \\ &= \frac{1}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} u \exp\left(-\frac{1}{2}u^2\right) du + \frac{\mu}{(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}u^2\right) du \end{aligned}$$

on making the substitution $u = (v - \mu)/\sigma$ in both integrands. The first integrand is an odd function, so the integral over \mathbb{R} is zero. The second term is μ by Example 7.1.21 and 7.1.22. Hence, $\mathbf{E}(X) = \mu$. \bullet

Expectation may be infinite, as the next example shows.

(6) **Example: Pareto Density** Let X have density $f(x) = (\alpha - 1)x^{-\alpha}$ for $x \geq 1$ and $\alpha > 1$. Then if $\alpha \leq 2$, the expected value of X is infinite because

$$\mathbf{E}(X) = \lim_{n \rightarrow \infty} \int_1^n \frac{(\alpha - 1)v}{v^\alpha} dv = (\alpha - 1) \lim_{n \rightarrow \infty} \int_1^n \frac{1}{v^{\alpha-1}} dv,$$

which diverges to ∞ for $\alpha - 1 \leq 1$. However, for $\alpha > 2$,

$$\mathbf{E}(X) = \int_1^{\infty} \frac{(\alpha - 1)}{v^{\alpha-1}} dv = \frac{(\alpha - 1)}{(\alpha - 2)}. \quad \bullet$$

Then again, the expectation of X may not exist, as the next example shows.

(7) **Example: Cauchy Density** Let X have density

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty.$$

Because $\int_0^a v(\pi(1+v^2))^{-1} dv$ diverges as $a \rightarrow -\infty$ and as $a \rightarrow +\infty$, X does not have an expected value. ●

It is appropriate to give a moment to considering why we define $\mathbf{E}(X)$ by Definition 1. This definition is at least plausible, by analogy with the definition

$$\mathbf{E}(X) = \sum_{v=-\infty}^{\infty} vf(v),$$

in the discrete case. Of course, Definition 1 is much more than just a plausible analogy, but a complete account of expectation is well beyond our scope. However, we can use Example 7.2.14 to give a little more justification for Definition 1. Let $k + r2^{-n} = a(k, r, n)$. Recall from (7.2.15) that $|S_n - X| < 2^{-n}$. Now by definition, because $S_n(X)$ is discrete,

$$\begin{aligned} \mathbf{E}(S_n(X)) &= \sum_{k,r} a(k, r, n) \int_{a(k,r,n)}^{a(k,r+1,n)} f(v)dv \\ &= \sum_{k,r} \int_{a(k,r,n)}^{a(k,r+1,n)} (vf(v) + (a(k, r, n) - v)f(v))dv. \end{aligned}$$

Because $|a(k, r, n) - v| < 2^{-n}$, it can be shown (with more work, which we omit) that

$$\mathbf{E}(S_n(X)) = \int_{-\infty}^{\infty} vf(v)dv + \epsilon_n = \mathbf{E}(X) + \epsilon_n,$$

where $\epsilon_n \rightarrow 0$ as $n \rightarrow \infty$. An explicit demonstration may be helpful here.

(8) **Example** Let X be uniform on $(0, 1)$ with mean value $\mathbf{E}(X) = \frac{1}{2}$. Then $S_n(X)$ is uniformly distributed on $\{0, 2^{-n}, 2 \cdot 2^{-n}, 3 \cdot 2^{-n}, \dots, 1 - 2^{-n}\}$. Therefore,

$$\begin{aligned} \mathbf{E}(S_n(X)) &= \sum_{r=0}^{2^n-1} r2^{-n} \cdot 2^{-n} = \frac{1}{2}(2^n - 1)2^{-n} \\ &= \mathbf{E}(X) - 2^{-(n+1)} \rightarrow \mathbf{E}(X), \quad \text{as } n \rightarrow \infty. \end{aligned} \quad \bullet$$

Thus, our definitions of expectation for discrete and continuous variables are at least consistent in some way. In more advanced books, a single definition of $\mathbf{E}(\cdot)$ is given, which is shown to yield our definitions as special cases.

Next we return to considering functions of random variables. Suppose we are given random variables Y and X related by $Y = g(X)$. What is $\mathbf{E}(Y)$?

If we know the density of X , then we may be able to find $\mathbf{E}(Y)$ by first discovering $f_Y(y)$, if it exists. This is often an unattractive procedure. We may do much better to use the following theorem, which we state without proof.

(9) **Theorem** Let random variables X and Y satisfy $Y = g(X)$, where X has density $f(x)$. Then Y has an expected value if $\int_{-\infty}^{\infty} |g(v)|f(v)dv < \infty$, and in this case,

$$(10) \quad \mathbf{E}(Y) = \int_{-\infty}^{\infty} g(v)f(v)dv.$$

The proof of this is straightforward but long. An heuristic discussion of the type above shows that if we represent the distribution of X as a limit of discrete distributions, and then formally proceed to this limit in Theorem 4.3.4, equation (10) is the result. Again, this only makes (10) plausible, it does not provide the proof, which is beyond our scope.

This important result implies that the useful consequences of Theorem 4.3.4 remain true for random variables with a density. In particular, Theorem 4.3.6 remains true; the proofs of most parts are just typographical variants of the proofs in the discrete case; just replace \sum by \int .

We describe one important and less trivial case in detail, namely, the analogy of Theorem 4.3.11.

(11) **Theorem: Tail integral** Let the nonnegative random variable X have density f , distribution F , and finite expected value $\mathbf{E}(X)$. Then

$$(12) \quad \mathbf{E}(X) = \int_0^{\infty} (1 - F(x))dx.$$

Proof For any finite y , we may integrate by parts to obtain

$$(13) \quad \int_0^y xf(x)dx = -x(1 - F(x))\Big|_0^y + \int_0^y (1 - F(x))dx.$$

But we have

$$y(1 - F(y)) = y \int_y^{\infty} f(x)dx \leq \int_y^{\infty} xf(x)dx \rightarrow 0$$

as $y \rightarrow \infty$, because $\mathbf{E}(X) < \infty$. Hence, we can let $y \rightarrow \infty$ in (13) to prove the theorem. ■

We can use this to prove a useful special case of Theorem 9.

(14) **Example** Let the nonnegative random variable X have density f , and let $g(X) \geq 0$. Show that $\mathbf{E}(g(X)) = \int_0^{\infty} g(v)f(v)dv$.

Solution

$$\begin{aligned} \mathbf{E}(g(X)) &= \int_0^{\infty} \mathbf{P}(g(X) \geq v) dv \quad \text{by (12)} \\ &= \int_0^{\infty} \int_{x:g(x) \geq v} f(x)dx dv = \int_0^{\infty} f(x) \int_0^{g(x)} dv dx = \int_0^{\infty} f(x)g(x)dx, \end{aligned}$$

as required. The interchange in the order of integration is justified by a theorem on double integrals, which we omit. ●

The various moments of a random variable with a density are defined just as they were for discrete random variables, that is to say: $\mu_k = \mathbf{E}(X^k)$, and

$$\sigma_k = \mathbf{E}((X - \mathbf{E}(X))^k).$$

(15) **Example: Normal Density** Let X have the density $N(0, \sigma^2)$. Find μ_k for all k .

Solution If k is odd, then $x^k \exp(-x^2/(2\sigma^2))$ is an odd function. Hence, $\mu_k = 0$ if k is odd. If $k = 2n$, then integrating by parts gives

$$\begin{aligned} \mu_{2n} &= \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \int_{-\infty}^{\infty} v^{2n} \exp(-v^2/(2\sigma^2)) dv \\ &= \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \left(-v^{2n-1} \sigma^2 \exp(-v^2/(2\sigma^2)) \Big|_{-\infty}^{\infty} \right. \\ &\quad \left. + \int_{-\infty}^{\infty} (2n-1) \sigma^2 v^{2n-2} \exp(-v^2/(2\sigma^2)) dv \right) \\ &= (2n-1) \sigma^2 \mu_{2n-2} = \sigma^{2n} \frac{(2n)!}{2^n n!} \end{aligned}$$

on iterating and observing that $\mu_0 = 1$. Hence, in particular, $\mu_2 = \sigma^2$. ●

Finally, and thankfully, we are pleased to record that the expectation $\mathbf{E}(X)$ of a continuous random variable X has the same useful basic properties that we established for the discrete case in Section 4.6. For convenience, we recall them here.

(16) **Theorem** Let a and b be constants, and let g and h be functions. Then:

(i) If $g(X)$ and $h(X)$ have finite mean, then

$$\mathbf{E}(g(X) + h(X)) = \mathbf{E}(g(X)) + \mathbf{E}(h(X)).$$

(ii) If $\mathbf{P}(a \leq X \leq b) = 1$, then $a \leq \mathbf{E}(X) \leq b$.

(iii) If h is nonnegative, then for $a > 0$, $\mathbf{P}(h(X) \geq a) \leq \mathbf{E}(h(X)/a)$.

(iv) **Jensen's inequality** If g is convex then $\mathbf{E}(g(X)) \geq g(\mathbf{E}(X))$.

Proof The proof is an exercise for you. When $h(x) = x^2$ in (iii), we have:

(17) **Chebyshev's inequality:** $\mathbf{P}(|X| \geq a) \leq \mathbf{E}X^2/a^2$. ■

7.5 Moment Generating Functions

In dealing with integer valued discrete random variables, we found the probability generating function exceptionally useful (see Chapter 6). It would be welcome to have such

a useful workhorse available for random variables with densities. Of course, if X has a density then $\mathbf{P}(X = x) = 0$, so we cannot expect the probability generating function to be of much use. Fortunately, another function will do the job.

(1) **Definition** If X has density f , then X has moment generating function $M_X(t)$ given by

$$(2) \quad M_X(t) = \mathbf{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tv} f(v) dv. \quad \blacktriangle$$

We are only interested in $M_X(t)$ for those values of t for which it is finite; this includes $t = 0$, of course. It is particularly pleasant when $M_X(t)$ exists in a neighbourhood of zero, but it is beyond our scope to explain all the reasons for this.

(3) **Example: Uniform Density** Let X be uniform on $[0, a]$. Find $\mathbf{E}(e^{tX})$. Where does it exist?

Solution

$$\mathbf{E}(e^{tX}) = \int_0^a \frac{1}{a} e^{tv} dv = \left[\frac{1}{at} e^{tv} \right]_0^a = \frac{e^{at} - 1}{at}.$$

This exists for all t , including $t = 0$, where it takes the value 1. ●

(4) **Example: Gamma Density** Recall from (7.1.24) that the gamma function $\Gamma(\alpha)$ is defined for any $\alpha > 0$ and $\lambda > 0$ by

$$(5) \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \lambda^{\alpha} e^{-\lambda x} dx.$$

Hence,

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0,$$

is the density of a random variable x . Find $\mathbf{E}(e^{tX})$. Where does it exist?

Solution

$$\mathbf{E}(e^{tX}) = \int_0^{\infty} e^{tv} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} v^{\alpha-1} e^{-\lambda v} dv = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} v^{\alpha-1} e^{-(\lambda-t)v} dv.$$

The integral exists if $\lambda > t$, and then making the substitution $(\lambda - t)v = u$ gives

$$(6) \quad M_X(t) = \left(\frac{\lambda}{\lambda - t} \right)^{\alpha} \int_0^{\infty} \frac{u^{\alpha-1}}{\Gamma(\alpha)} e^{-u} du = \left(\frac{\lambda}{\lambda - t} \right)^{\alpha} \quad \text{by (5),}$$

for $-\infty < t < \lambda$. ●

(7) **Example: Normal Density** Let X be a standard normal random variable. Then

$$\begin{aligned}\sqrt{2\pi}M_X(t) &= \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}x^2 + tx\right) dx = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(x-t)^2 + \frac{1}{2}t^2\right) dx \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}v^2\right) dv, \quad \text{setting } x-t=v, \\ &= e^{\frac{1}{2}t^2} \sqrt{2\pi}.\end{aligned}$$

So $M_X(t) = e^{\frac{1}{2}t^2}$. Now by (7.2.4) if Y is $N(\mu, \sigma^2)$, $M_Y(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. ●

You may ask, why is $M_X(t)$ called the moment generating function? The answer lies in the following formal expansion.

$$(8) \quad \mathbf{E}(e^{tX}) = \mathbf{E}\left(\sum_{k=0}^{\infty} \frac{X^k t^k}{k!}\right) = \sum_{k=0}^{\infty} \mathbf{E}\frac{(X^k)t^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k t^k}{k!}.$$

Thus, provided the interchange of expectation and summation at (8) is justified, we see that $M_X(t)$ is the (exponential) generating function of the moments μ_k .

Note that the word “exponential” is always omitted in this context, and that the required interchange at (8) is permissible if $M_X(t)$ exists in an interval that includes the origin.

You may also ask, do we always know the density $f_X(x)$, if we know $M_X(t)$? After all, the probability generating function uniquely determines the corresponding mass function. Unfortunately, the answer is no in general because densities not uniquely determined by their moments do exist. However, none appear here; every density in this book is uniquely determined by its moment generating function (if it has one). We state the following *inversion theorem* without proof.

(9) **Theorem** If X has moment generating function $M(t)$, where for some $a > 0$, $M(t) < \infty$ for $|t| < a$, then the distribution of X is determined uniquely.

$$\text{Furthermore, } M(t) = \sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbf{E}(X^k).$$

The moment generating function is especially useful in dealing with sequences of random variables; the following theorem is the basis of this assertion. We state it without proof.

(10) **Theorem: Continuity Theorem** Let $(F_n(x); n \geq 1)$ be a sequence of distribution functions with corresponding moment generating functions $(M_n(t); n \geq 1)$ that exist for $|t| < b$. Suppose that as $n \rightarrow \infty$ $M_n(t) \rightarrow M(t)$ for $|t| \leq a < b$, where $M(t)$ is the m.g.f. of the distribution $F(x)$. Then, as $n \rightarrow \infty$, $F_n(x) \rightarrow F(x)$ at each point x where $F(x)$ is continuous.

The main application of this theorem arises when $M(t) = e^{\frac{1}{2}t^2}$ and $F(x) = \Phi(x)$, as we see in Chapter 8 when we come to the celebrated central limit theorem. Here is a preliminary note.

Note: The O - o Notation In considering limits of sequences of functions, we quite often produce large and unwieldy expressions of which only one or two terms remain in the limit. Rather than keep a precise record of the essentially irrelevant terms, it is convenient to have a special compact notation for them.

Definition If $g(n)$ and $h(n)$ are two functions of n , then we write

$$h(n) = O(g(n)) \quad \text{as } n \rightarrow \infty$$

if $|h(n)/g(n)| < c$ for all large enough n and some finite constant c . ▲

For example, as $n \rightarrow \infty$,

$$n^2 + \log n = O(n^2) \quad \text{with } c = 2$$

and

$$n^2 + n^{\frac{3}{2}} = O(n^2) \quad \text{with } c = 2.$$

Observe that this is an abuse of notation (= being the abused symbol) because it does not follow from these two examples that $\log n = n^{\frac{3}{2}}$. Also, if $h(n) = O(g(n))$ and $k(n) = O(g(n))$, then $h(n) + k(n) = O(g(n))$.

A similar definition holds for small values of the argument.

Definition If $g(x)$ and $h(x)$ are two functions of x , then we write $h(x) = O(g(x))$ as $x \rightarrow 0$, if $|h(x)/g(x)| < c$ for all small enough x and some constant c . ▲

Often, an even cruder representation will suffice.

Definition If $g(x)$ and $h(x)$ are two functions of x , then we write $h(x) = o(g(x))$ as $x \rightarrow \infty$, if $\lim_{x \rightarrow \infty} (h(x)/g(x)) = 0$. Likewise, $h(x) = o(g(x))$ as $x \rightarrow 0$ if $\lim_{x \rightarrow 0} (h(x)/g(x)) = 0$. ▲

For example, $x^2 = o(x)$ as $x \rightarrow 0$ and $x = o(x^2)$ as $x \rightarrow \infty$. For another example, $x + x \log x + x^2 = o(1)$ as $x \rightarrow 0$.

We use this new notation in the following famous result.

(11) **Example: de Moivre–Laplace Theorem** For each $n \geq 1$, let X_n be a binomial random variable with parameters n and p . Let $q = 1 - p$, and define

$$Y_n = \frac{X_n - np}{(npq)^{\frac{1}{2}}}.$$

Show that as $n \rightarrow \infty$

$$\mathbf{P}(Y_n \leq x) \rightarrow \Phi(x) = \int_{-\infty}^x (2\pi)^{-\frac{1}{2}} e^{-y^2/2} dy.$$

Solution We use Theorem 10. First calculate the moment generating function

$$(12) \quad \mathbf{E}(e^{tY_n}) = \mathbf{E} \left(\exp \left(\frac{t(X_n - np)}{(npq)^{\frac{1}{2}}} \right) \right) = \left[\mathbf{E} \exp \left(\frac{t(X_1 - p)}{(npq)^{\frac{1}{2}}} \right) \right]^n \\ = \left[p \exp \left(\frac{qt}{(npq)^{\frac{1}{2}}} \right) + q \exp \left(\frac{-pt}{(npq)^{\frac{1}{2}}} \right) \right]^n.$$

Next we expand the two exponential terms in (12) to give

$$(13) \quad \mathbf{E}(e^{tY_n}) = \left(1 + \frac{t^2}{2n} + O \left(n^{-\frac{3}{2}} \right) \right)^n.$$

Now we recall the useful result that says that, for constant a ,

$$(14) \quad \lim_{n \rightarrow \infty} \left(1 + \frac{a}{n} + o(n^{-1}) \right)^n = e^a.$$

Applying this to (13) shows that

$$\lim_{n \rightarrow \infty} \mathbf{E}(e^{tY_n}) = e^{\frac{1}{2}t^2},$$

which is the m.g.f. of the standard normal distribution, as required. [More demanding readers should note that they can prove (14) by first taking logarithms.] ●

The appearance of the normal distribution in these circumstances is one of the most remarkable results in the theory of probability. The first proof, due to de Moivre, was greatly improved by Laplace. Their methods were different from those used here, relying on fairly precise direct estimates of the binomial probabilities. We outline a modern version of their proof in Example 7.20.

7.6 Conditional Distributions

Just as in the discrete case, it is often necessary to consider the distribution of a random variable X conditional upon the occurrence of some event A . By definition of conditional probability, we have

$$(1) \quad F_{X|A}(x) = \mathbf{P}(X \leq x|A) = \mathbf{P}(\{\omega: X \leq x\} \cap A) / \mathbf{P}(A) \\ = \mathbf{P}(X \leq x; A) / \mathbf{P}(A), \text{ say.}$$

(Obviously, A has to be in \mathcal{F} , the event space.) The case that arises most commonly is when A is an event of the form

$$(2) \quad A = \{\omega: a < X \leq b\};$$

that is, we seek the distribution of X conditional on its lying in some subset of its range.

(3) Example Let $a < b < c < d$. Let X be uniform on (a, d) , and let $A = \{\omega: b < X(\omega) \leq c\}$.

Then, by (1),

$$\begin{aligned} \mathbf{P}(X \leq x|A) &= \frac{\mathbf{P}(X \leq x; b < X \leq c)}{\mathbf{P}(b < X \leq c)} = \begin{cases} \frac{x-b}{d-a} / \left(\frac{c-b}{d-a} \right) & \text{if } b < x \leq c \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{x-b}{c-b} \quad \text{for } b < x \leq c. \end{aligned}$$

Thus, the distribution of X given A is just uniform on (b, c) . More generally, it is easy to see that a uniform random variable, constrained to lie in any subset A of its range, is uniformly distributed over the subset A . ●

Because $\mathbf{P}(X \leq x|A)$ is a distribution, it may have an expectation. For example, suppose that X has density f , and A is given by (2). Then, by (1),

$$(4) \quad \mathbf{P}(X \leq x|A) = \frac{F(x) - F(a)}{F(b) - F(a)} \quad \text{for } a < x \leq b,$$

and differentiating yields the conditional density

$$(5) \quad f_{X|A}(x) = \begin{cases} \frac{f(x)}{F(b) - F(a)} & \text{if } a < x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $\int_a^b f_{X|A}(v)dv = 1$, as it must. Then we may define the *conditional expectation*

$$(6) \quad \begin{aligned} \mathbf{E}(X|A) &= \int_a^b \frac{vf(v)}{F(b) - F(a)} dv \\ &= a + \int_a^b \frac{F(b) - F(v)}{F(b) - F(a)} dv, \quad \text{on integrating by parts,} \\ &= \int_0^\infty (1 - F_{X|A}(v))dv \end{aligned}$$

on using (4). Notice that this is in agreement with Theorem 7.4.11, as of course it must be.

(7) **Example: Exponential Density and Lack-of-Memory** Let X be exponentially distributed with parameter λ . Show that

$$(8) \quad \mathbf{P}(X > s + t|X > s) = e^{-\lambda t} = \mathbf{P}(X > t).$$

Find $\mathbf{E}(X|X > s)$ and $\mathbf{E}(X|X \leq s)$.

Solution Trivially,

$$\mathbf{P}(X > s + t|X > s) = \mathbf{P}(X > s + t)/\mathbf{P}(X > s) = e^{-\lambda(s+t)}/e^{-\lambda s} = e^{-\lambda t}.$$

Hence,

$$(9) \quad \mathbf{E}(X|X > s) = s + \int_0^\infty e^{-\lambda t} dt = s + \mathbf{E}(X).$$

We remark that the remarkable identity (8) is known as the *lack-of-memory property* of the exponential distribution. Finally,

$$\mathbf{E}(X|X \leq s) = \int_0^s \frac{\mathbf{P}(s \geq X > v)}{\mathbf{P}(s \geq X)} dv = \frac{1}{\lambda} - \frac{s}{e^{\lambda s} - 1}. \quad \bullet$$

7.7 Ageing and Survival

Many classic examples of continuous random variables arise as waiting times or survival times. For instance, the time until the cathode-ray tube in your television fails, the time until you are bitten by a mosquito after disembarking in the tropics, the time until a stressed metal component fails due to fatigue. For definiteness, we consider the *lifetime* T of some *device* or component. The device is said to *fail* at time T .

It is often useful to quantify the ageing process of a device; in particular, we may want to compare a device of given age with a new one. (We are all familiar with the fact that it is not necessarily always a good thing to replace a working component with a new one. This fact is embodied in the popular saying: “If it works, don’t fix it”) Let T have distribution F and density f . The following quantities turn out to be of paramount importance in comparing devices of different ages.

The *survival function*

$$(1) \quad \bar{F}(t) = 1 - F(t) = \mathbf{P}(T > t).$$

The *hazard function*

$$(2) \quad H(t) = -\log(1 - F(t)).$$

The *hazard rate function*

$$(3) \quad r(t) = \frac{f(t)}{\bar{F}(t)} = \frac{f(t)}{1 - F(t)} = \frac{dH(t)}{dt}.$$

The last equality explains why $r(t)$ is called the hazard rate. Integrating (3) yields

$$(4) \quad \exp\left(-\int_0^t r(s) ds\right) = \bar{F}(t).$$

Before we explain the significance of these quantities, you are warned that terminology in the literature of ageing is quite chaotic. Note that:

- (i) The survival function is also known as the survivor function, reliability function, or hazard function.
- (ii) The hazard function is also known as the log-survivor function.
- (iii) The hazard rate function is also known as the failure rate function, mortality function, or hazard function.

Beware!

Now let A_t denote the event that $T > t$. Then

$$F_{T|A_t}(s+t) = \mathbf{P}(T \leq s+t | T > t) = \frac{F(s+t) - F(t)}{1 - F(t)}.$$

This is the probability that the device fails during $(t, t+s)$, given that it has not failed by time t . Now

$$(5) \quad \lim_{s \rightarrow 0} \frac{1}{s} F_{T|A_t}(s+t) = (1 - F(t))^{-1} \lim_{s \rightarrow 0} \frac{F(t+s) - F(t)}{s} = \frac{f(t)}{1 - F(t)} = r(t).$$

Thus, $r(t)$ may be thought of as the “intensity” of the probability that a device aged t will fail.

- (6) **Example: Exponential Life** If T has an exponential density, then $\bar{F}(t) = e^{-\lambda t}$, $H(t) = \lambda t$, and $r(t) = \lambda$. This constant hazard rate is consonant with the lack-of-memory property mentioned in Example 7.6.7. Roughly speaking, the device cannot remember how old it is, and so the failure intensity remains constant. ●

We see that intuitively there is a distinction between devices for which $r(t)$ increases, essentially they are “wearing out,” and those for which $r(t)$ decreases, they are “bedding in.”

A simple and popular density in this context is the Weibull density, which can exhibit both types of behaviour.

- (7) **Example: Weibull Life** If T has density

$$f(t) = \alpha t^{\alpha-1} \exp(-t^\alpha), \quad t > 0, \alpha > 0,$$

then it has distribution $F(t) = 1 - \exp(-t^\alpha)$. Hence, $\bar{F}(t) = \exp(-t^\alpha)$, and so

$$\frac{\mathbf{P}(T > t+s | T > s)}{\mathbf{P}(T > t)} = \exp(-(t+s)^\alpha + s^\alpha + t^\alpha),$$

which is > 1 or < 1 according as $\alpha < 1$ or $\alpha > 1$. [To see this, just consider the stationary value of $x^\alpha + (1-x)^\alpha - 1$ at $x = \frac{1}{2}$.] Hence, if $\alpha < 1$, the chance of lasting a further time t (conditional on $T > s$) increases with s . However, if $\alpha > 1$, this chance decreases with s . ●

The behaviour of $r(t)$ is not the only measure of comparison between new and old devices. There is a large hierarchy of measures of comparison, which we display formally as follows [in the notation of (1)–(3)].

- (8) **Definition**

- (i) If $r(t)$ increases, then T is (or has) increasing failure rate, denoted by IFR.
(ii) If $\frac{H(t)}{t}$ increases, then T is (or has) increasing failure rate average, denoted by IFRA.
(iii) If for all $s \geq 0, t \geq 0$,

$$H(s+t) \geq H(s) + H(t),$$

then T is new better than used, denoted by NBU.

(iv) If for all $t \geq 0$

$$\mathbf{E}(T) \geq \mathbf{E}(T - t|A_t),$$

then T is new better than used in expectation, denoted by NBUE.

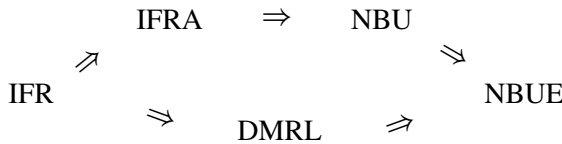
(v) If for all $0 \leq s < t < \infty$

$$\mathbf{E}(T - s|A_s) \geq \mathbf{E}(T - t|A_t),$$

then T has (or is) decreasing mean residual life, denoted by DMRL. ▲

The random variable T may also be decreasing failure rate (DFR), decreasing failure rate on average (DFRA), new worse than used (NWU), new worse than used in expectation (NWUE), or increasing mean residual life (IMRL). All these are defined in the obvious way, analogous to (i)–(v).

It can be shown that the following relationships hold between these concepts:



Some of these implications are trivial, and some are established in Example 7.17 below. These ideas are linked to another concept, that of stochastic ordering.

7.8 Stochastic Ordering

As in Section 7.7, let T be a nonnegative random variable. In general, let $R(s)$ be a random variable whose distribution is that of $T - s$ given that $T > s$, namely,

$$(1) \quad F_R(x) = \mathbf{P}(T - s \leq x | T > s).$$

We refer to $R(s)$ as the *residual life* (of T at s).

The above example shows that if T has the exponential density, then its residual life is also exponentially distributed with constant mean.

More generally, $F_{R(s)}$ may depend on s , and more significantly it may do so in a systematic way; the following definition is relevant here.

(2) **Definition** Let X and Y be random variables. If

$$(3) \quad \bar{F}_X(x) \geq \bar{F}_Y(x) \quad \text{for all } x,$$

then X is said to be stochastically larger than Y . ▲

Now we can supply a connection with the ideas of the preceding section (7.7).

(4) **Example** If T is a random variable with residual life $R(s)$, $s > 0$, show that T has increasing failure rate if and only if $R(s)$ is stochastically larger than $R(t)$ for all $s < t$.

Solution First, we find

$$\begin{aligned}
 (5) \quad \mathbf{P}(R(t) > x) &= \mathbf{P}(T - t > x | T > t) = \bar{F}(t+x)/\bar{F}(t) \\
 &= \exp\left(-\int_0^{t+x} r(s)ds\right) \exp\left(\int_0^t r(s)ds\right) \quad \text{by (7.7.4)} \\
 &= \exp\left(-\int_0^{t+x} r(s)ds\right).
 \end{aligned}$$

Differentiating (5) with respect to t , we have

$$(6) \quad \frac{\partial}{\partial t} \mathbf{P}(R(t) > x) = (r(t) - r(t+x)) \exp\left(-\int_t^{t+x} r(s)ds\right).$$

Because $\exp(-\int r ds)$ is positive, and $r(t) - r(t+x)$ is positive or negative according as T is DFR or IFR, the result follows. ●

Finally, we have the useful:

(7) **Theorem** *If X is stochastically larger than Y , then $\mathbf{E}(X) \geq \mathbf{E}(Y)$.*

Proof We prove this when $X \geq 0$ and $Y \geq 0$. (The general result is left as an exercise.) From Theorem 7.4.11,

$$\begin{aligned}
 \mathbf{E}(X) &= \int_0^\infty \bar{F}_X(x)dx \geq \int_0^\infty \bar{F}_Y(x)dx \quad \text{by hypothesis,} \\
 &= \mathbf{E}(Y). \quad \blacksquare
 \end{aligned}$$

7.9 Random Points

Picking a point Q at random in the interval $(0, 1)$ yielded the uniform density (of the length OQ). It is intuitively attractive to consider problems that involve picking one or more points at random in other nice geometric figures, such as discs, squares, triangles, spheres, and so on. Indeed this idea is so natural that mathematicians had already started doing this kind of thing in the eighteenth century, and one of the most celebrated articles on the subject is that of M.W. Crofton in the 1885 edition of the *Encyclopaedia Britannica*. Such questions also have applications in statistics.

Confining ourselves to two dimensions for definiteness, suppose a point Q is picked at random in a region R of area $|R|$. Then it is natural to let the probability $\mathbf{P}(S)$, that Q lies in a set $S \subseteq R$, be given by

$$(1) \quad \mathbf{P}(S) = \frac{|S|}{|R|},$$

where, now, $|S|$ denotes the area of S . It follows from the properties of area that $\mathbf{P}(\cdot)$ has the required properties of a probability function, and we can proceed to solve various simple problems using elementary geometry. The following is typical.

(2) **Example** A point Q is picked at random in the unit square. What is the probability ν that it is nearer to the centre O of the square than to its perimeter?

Solution By symmetry, we need to consider only the sector $0 \leq y \leq x \leq \frac{1}{2}$. Then the point (x, y) is nearer to O than the perimeter if $\sqrt{x^2 + y^2} < \frac{1}{2} - x$; that is, if in this sector,

$$x < \frac{1}{4} - y^2, \text{ for } 0 \leq y \leq \frac{1}{2}(\sqrt{2} - 1).$$

Hence, the area is given by an integral and

$$v = 8 \int_0^{\frac{1}{2}(\sqrt{2}-1)} \left(\frac{1}{4} - y^2 - y \right) dy = \frac{4}{3}\sqrt{2} - \frac{5}{3}. \quad \bullet$$

An equally trivial but much more important example is the following.

- (3) **Example** Let $f = f(x)$ be an integrable function with $0 \leq f(x) \leq 1$ for $0 \leq x \leq 1$. Let Q be picked at random in the unit square, and let A_v be the set of points such that $0 \leq x \leq v$ and $0 \leq y \leq f(x)$. Then, from (1),

$$(4) \quad \mathbf{P}(Q \in A_v) = \int_0^v f(x) dx. \quad \bullet$$

This trivial result has at least two important applications. The first we have met already in Example 5.8.9.

- (5) **Example: Hit-or-Miss Monte Carlo Integration** Let $f(x)$ and Q be as defined in Example 3, and declare Q a *hit* if Q lies below $f(x)$, $0 \leq x \leq 1$. Then the probability of a hit is $\mathbf{P}(A_1) = \int_0^1 f(x) dx$. Now we pick a sequence of such points Q_1, Q_2, \dots and let X_n be the number of hits. If points are picked independently, then X_n is a binomial random variable with parameters n and $\mathbf{P}(A_1)$, and we have shown that as $n \rightarrow \infty$, for $\epsilon > 0$,

$$\mathbf{P} \left(|n^{-1} X_n - \int_0^1 f(x) dx| > \epsilon \right) \rightarrow 0.$$

This therefore offers a method for evaluating the integral $\int_0^1 f(x) dx$. In practice, one would be unlikely to use this method in one dimension, but you might well use the analogous method to evaluate $\int f(\mathbf{x}) d\mathbf{x}$, where \mathbf{x} is a vector in (say) 11 dimensions. \bullet

- (6) **Example: Simulation** With Q and $f(x)$ defined as above, consider the probability that Q lies in A_v given that it is a hit. By definition, this has probability

$$\mathbf{P}(A_v | A_1) = \frac{\int_0^v f(x) dx}{\int_0^1 f(x) dx}.$$

By inspection, the function $F(v) = \mathbf{P}(A_v | A_1)$ is the distribution function of the x -coordinate of Q given that it is a hit. This procedure therefore offers a method of simulating

a random variable X with density function

$$(7) \quad f_X(x) = \frac{f(x)}{\int_0^1 f(x) dx}.$$

You can just pick a point Q and, if it is a hit, let its x -coordinate be X . ●

A natural next step is to consider events defined jointly by a number of points picked independently in a region R . One famous example is Sylvester's problem: for four points picked at random in R , what is the probability that one of them lies in the triangle formed by the other three? This is too difficult for us, but we can consider an amusing simpler problem to illustrate a few of the basic ideas.

- (8) **Example: Two Points in a Disc** Let $\lambda(r)$ be the expected value of the distance $L(r)$ between two points Q_1 and Q_2 , each distributed uniformly (and independently) over a disc of radius r . Show that

$$(9) \quad \lambda(r) = \frac{128r}{45\pi}.$$

Solution This can be done by a brutal integration; here is a better way, discovered by M.V. Crofton in 1885.

Consider a disc of radius $x+h$, which we may think of as a disc D of radius x , surrounded by an annulus A of width h . Then, if Q_1 and Q_2 are dropped at random on to the disc of radius $x+h$, we have (using independence and the properties of the uniform density) that

$$(10) \quad \mathbf{P}(Q_1 \in D \cap Q_2 \in D) = \left(\frac{\pi x^2}{\pi(x+h)^2} \right)^2 = 1 - \frac{4h}{x} + o(h).$$

Also,

$$\mathbf{P}(Q_1 \in D \cap Q_2 \in A) = \frac{\pi x^2}{\pi(x+h)^2} \left(1 - \frac{\pi x^2}{\pi(x+h)^2} \right) = \frac{2h}{x} + o(h)$$

and $\mathbf{P}(Q_1 \in A \cap Q_2 \in A) = o(h)$. Hence, by conditional expectation,

$$(11) \quad \begin{aligned} \lambda(x+h) &= \mathbf{E}(L(x+h) | Q_1 \in D; Q_2 \in D) \left(1 - \frac{4h}{x} + o(h) \right) \\ &\quad + 2\mathbf{E}(L(x+h) | Q_1 \in D; Q_2 \in A) \left(\frac{2h}{x} + o(h) \right) + o(h). \end{aligned}$$

Now $\mathbf{E}(L(x+h) | Q_1 \in D; Q_2 \in A)$ is just the mean distance of a random point Q_1 in a disc of radius x , from a point Q_2 on its circumference (plus a quantity that is $o(h)$). Hence, taking plane polar coordinates with Q_2 as origin:

$$\begin{aligned} \mathbf{E}(L(x+h) | Q_1 \in D; Q_2 \in A) &= \frac{1}{\pi x^2} \int_{-\pi/2}^{\pi/2} \int_0^{2x \cos \theta} v^2 dv d\theta + o(h) \\ &= \frac{32x}{9\pi} + o(h). \end{aligned}$$

Returning to (11), note that $\mathbf{E}(L(x+h)|Q_1 \in D; Q_2 \in D) = \lambda(x)$; hence, rearranging (11) and letting $h \rightarrow 0$ gives

$$\frac{d\lambda(x)}{dx} = \lim_{h \rightarrow 0} \frac{1}{h} \left(\lambda(x+h) - \lambda(x) \right) = \frac{-4}{x} \lambda(x) + \frac{128}{9\pi}.$$

Integrating this, and observing that $\lambda(0) = 0$, we have

$$\lambda(x) = \frac{128x}{45\pi}.$$

Using the same idea, and with a lot more toil, we can find the density of L .

The next natural step is to pick lines (or other objects) at random and ask how they divide up the region R in random tessellations or coverings. This is well beyond our scope, but the trivial Example 7.18 illustrates some of the problems.

7.10 Review and Checklist for Chapter 7

We introduced the class of random variables having a density $f_X(x)$ and a distribution $F_X(x)$. These take one of an uncountable number of values in \mathbb{R} and are called “absolutely continuous.” The familiar ideas of expectation, conditioning, functions, and generating functions are explored in this new context, together with some applications.

SYNOPSIS OF FORMULAE:

Key Rule: $\mathbf{P}(X \in B) = \int_{x \in B} f_X(x) dx.$

Distribution and density:

$$F_X(x) = \int_{-\infty}^x f_X(y) dy = \mathbf{P}(X \leq x).$$

For small h , $\mathbf{P}(x < X \leq x+h) \simeq f_X(x)h$, and if $F(x)$ is differentiable $\frac{dF}{dx} = f(x)$.
 $F(x)$ is nondecreasing; $\lim_{x \rightarrow -\infty} F(x) = 0$;

$$\lim_{x \rightarrow +\infty} F(x) = \int_{-\infty}^{\infty} f(u) du = 1, \quad \text{if } X \text{ is proper.}$$

Mixture: If f and g are densities, then so is

$$h = \lambda f + (1 - \lambda)g, \quad 0 \leq \lambda \leq 1.$$

Functions: If continuous random variables X and Y are such that $Y = g(X)$ for some function $g(\cdot)$ that is differentiable and strictly increasing, then

$$f_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} [g^{-1}(y)],$$

where $g^{-1}(\cdot)$ is the inverse function of g . In general, we can write

$$f_Y(y) = \frac{d}{dy} \int_{x:g(x) \leq y} f_X(x) dx,$$

and proceed by ad hoc arguments.

Expectation: A random variable X has an expected value $\mathbf{E}X$ provided that $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$, and then

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

When $X > 0$,

$$\mathbf{E}X = \int_0^{\infty} \mathbf{P}(X > x) dx.$$

If random variables X and Y are such that $Y = g(X)$ and X is continuous, then Y has an expected value if $\int_{-\infty}^{\infty} |g(x)| f_X(x) dx < \infty$ and

$$\mathbf{E}Y = \mathbf{E}g(X) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Moments: In particular, if $g(X) = X^r$, this yields the r th moment μ_r of X . When $X > 0$,

$$\mathbf{E}X^r = \int_0^{\infty} r x^{r-1} \mathbf{P}(X > x) dx.$$

When $g(X) = \exp(tX)$, this yields the m.g.f.,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx.$$

Conditioning: Any event B in Ω may condition a random variable X on Ω , leading to a conditional distribution and density,

$$F_{X|B}(x|B) = \mathbf{P}(X \leq x|B)$$

and

$$f_{X|B}(x|B) = \frac{d}{dx} F_{X|B}(x|B),$$

when the derivative exists, with the

Key Rule:

$$\mathbf{P}(X \in A|B) = \int_{x \in A} f(x|B) dx.$$

Such conditioned random variables may have an expectation if

$$\int_{-\infty}^{\infty} |x| f_{X|B}(x|B) dx < \infty,$$

Table 7.1. Continuous random variables and their associated characteristics

X	$f(x)$	$\mathbf{E}X$	$\text{var } X$	m.g.f.
Uniform	$(b-a)^{-1}, a \leq x \leq b$	$\frac{1}{2}(b+a)$	$\frac{1}{12}(b-a)^2$	$\frac{e^{bt}-e^{at}}{t(b-a)}$
Exponential	$\lambda e^{-\lambda x}, x \geq 0$	λ^{-1}	λ^{-2}	$\lambda/(\lambda-t)$
Normal $N(\mu, \sigma^2)$	$(2\pi)^{-1/2} \sigma^{-1} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},$ $-\infty < x < \infty$	μ	σ^2	$\exp(\mu t + \frac{1}{2}\sigma^2 t^2)$
Gamma	$x^{r-1} e^{-\lambda x} \frac{\lambda^r}{(\sigma-1)!}, x \geq 0$	$r\lambda^{-1}$	$r\lambda^{-2}$	$\left(\frac{\lambda}{\lambda-t}\right)^r$
Laplace	$\frac{1}{2}\lambda \exp(-\lambda x)$	0	$2\lambda^{-2}$	$\frac{\lambda^2}{\lambda^2-t^2}$
Cauchy	$\{\pi(1+x^2)\}^{-1}$			

and then

$$\mathbf{E}(X|B) = \int_{-\infty}^{\infty} x f_{X|B}(x|B) dx.$$

Table 7.1 gives some useful continuous random variables with their elementary properties.

Checklist of Terms for Chapter 7

- 7.1 distribution function
 - density
 - standard normal density $\phi(x)$
 - mixture
- 7.2 functions
 - inverse function
- 7.3 simulation
 - composition
- 7.4 expected value
 - expectation of functions
 - tail integral for expectation
- 7.5 moment generating function
 - continuity theorem
 - O - o notation
 - de Moivre–Laplace theorem
- 7.6 conditional distribution
 - conditional density
 - conditional expectation
 - lack-of-memory property
- 7.7 hazard function
 - hazard rate
- 7.8 stochastically larger
- 7.9 geometrical probability

WORKED EXAMPLES AND EXERCISES

7.11 Example: Using a Uniform Random Variable

The random variable U is uniformly distributed on $(0, 1)$.

(a) Can you use U to get a random variable with density

$$(1) \quad f_0(y) = 12 \left(y - \frac{1}{2} \right)^2 \quad \text{for } 0 < y < 1?$$

(b) Actually, you really want a random variable with density

$$(2) \quad f(x) = 3 \left(\left(x - \frac{1}{2} \right)^2 + \frac{1}{8} |1 - 2x|^{\frac{1}{2}} \right) \quad \text{for } 0 < x < 1,$$

and in your pocket is a fair coin. Explain how the coin is useful.

Solution If $g(U)$ is a continuous increasing function, and $Y = g(U)$, then

$$F_Y(y) = \mathbf{P}(g(U) \leq y) = \mathbf{P}(U \leq g^{-1}(y)) = g^{-1}(y) \quad \text{because } U \text{ is uniform.}$$

From (1), we have the distribution of interest

$$F_Y(y) = \int_0^y 12 \left(y - \frac{1}{2} \right)^2 dy = 4 \left(y - \frac{1}{2} \right)^3 + \frac{1}{2}.$$

Hence, if we find a function $g(\cdot)$ such that

$$g^{-1}(y) = 4 \left(y - \frac{1}{2} \right)^3 + \frac{1}{2},$$

then $g(U)$ has the density (1) as required. Setting $y = g(u)$ and solving

$$u = 4 \left(g(u) - \frac{1}{2} \right)^3 + \frac{1}{2},$$

we find immediately that

$$g(u) = \left(\frac{u}{4} - \frac{1}{8} \right)^{\frac{1}{3}} + \frac{1}{2}$$

is the required function $g(\cdot)$.

For the second part, we notice that

$$\frac{1}{|1 - 2x|^{\frac{1}{2}}} = \frac{1}{(1 - 2x)^{\frac{1}{2}}} \quad \text{if } 0 < x < \frac{1}{2},$$

and that

$$f_1(x) = \begin{cases} (1 - 2x)^{-\frac{1}{2}} & \text{if } 0 < x < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases}$$

is a density function. By the method of the first part or by inspection, we see that

$$g_1(U) = \frac{1}{2}(1 - U^2)$$

is a random variable with density $f_1(x)$. (To see this, just make the simple calculation

$$\mathbf{P}(g_1(U) \leq x) = \mathbf{P}\left(\frac{1}{2}(1 - U^2) \leq x\right) = \mathbf{P}(U \geq (1 - 2x)^{\frac{1}{2}}) = 1 - (1 - 2x)^{\frac{1}{2}},$$

and differentiate to get the density f_1 .)

Likewise,

$$f_2(x) = \begin{cases} (2x - 1)^{-\frac{1}{2}} = |1 - 2x|^{\frac{1}{2}} & \text{if } \frac{1}{2} < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

is a density function, and

$$g_2(U) = \frac{1}{2}(1 + U^2)$$

is a random variable with density f_2 . (You can check this, as we did for g_1 and f_1 .)

Now you take the coin and toss it three times. Let A be the event that you get either three heads or three tails, B the event that you get two heads and a tail (in any order) and C the event that you get two tails and a head (in any order). Define the random variable

$$(3) \quad X = \begin{cases} \left(\frac{U}{4} - \frac{1}{8}\right)^{\frac{1}{3}} + \frac{1}{2} & \text{if } A \text{ occurs} \\ \frac{1}{2}(1 + U^2) & \text{if } B \text{ occurs} \\ \frac{1}{2}(1 - U^2) & \text{if } C \text{ occurs.} \end{cases}$$

Then the density of X is just a mixture of the densities of $g(U)$, $g_1(U)$, and $g_2(U)$, namely,

$$f_X(x) = \frac{1}{4}f_0(x) + \frac{3}{8}f_1(x) + \frac{3}{8}f_2(x) = 3\left(x - \frac{1}{2}\right)^2 + \frac{3}{8}|1 - 2x|^{-\frac{1}{2}},$$

as required. [We defined a mixture in (7.1.29).]

- (4) **Exercise** Explain how you would use U to get a random variable with density

$$f(x) = \frac{3}{4}\left(1 + (2x - 1)^2\right) \quad \text{if } 0 < x < 1.$$

- (5) **Exercise** Show that $Y = \gamma(-\log U)^{\frac{1}{\beta}}$ has a *Weibull* distribution.
 (6) **Exercise** Let the random variable X be defined by

$$X = \begin{cases} (2U)^{\frac{1}{2}} & \text{if } U < \frac{1}{2} \\ 2 - (2 - 2U)^{\frac{1}{2}} & \text{if } U \geq \frac{1}{2}. \end{cases}$$

Show that X has a triangular density on $[0, 2]$.

- (7) **Exercise** Find the densities of:
 (a) $\tan(\pi U)$.
 (b) $\tan(\frac{\pi}{2}U)$.

7.12 Example: Normal Distribution

Let

$$(1) \quad \phi(x) = (2\pi)^{-\frac{1}{2}} e^{-x^2/2}; \Phi(x) = \int_{-\infty}^x \phi(u) du.$$

(a) Define the sequence of functions $H_n(x); n \leq 0$, by

$$(2) \quad (-)^n \frac{d^n \phi(x)}{dx^n} = H_n(x) \phi(x); H_0 = 1.$$

Show that $H_n(x)$ is a polynomial in x of degree n . What is $H_1(x)$?

(b) Define Mills' ratio $r(x)$ by $r(x)\phi(x) = 1 - \Phi(x)$. Show that for $x > 0$

$$(3) \quad \frac{1}{x} - \frac{1}{x^3} < r(x) < \frac{1}{x}.$$

Solution First, make the important observation that

$$(4) \quad \frac{d\phi}{dx} = \frac{d}{dx}((2\pi)^{-\frac{1}{2}} e^{-x^2/2}) = -x\phi.$$

(a) We use induction. First, by (2),

$$-H_{n+1}(x)\phi(x) = +\frac{d}{dx}(H_n(x)\phi(x)) = H'_n(x)\phi(x) - H_n(x)x\phi(x), \quad \text{by (4).}$$

Hence,

$$(5) \quad H_{n+1}(x) = xH_n(x) - H'_n(x)$$

and, by (2) and (4), $H_1(x) = x$. The result follows by induction as claimed.

(b) For the right-hand inequality, we consider

$$\begin{aligned} 1 - \Phi(x) &= \int_x^\infty \phi(u) du \leq \int_x^\infty \frac{u}{x} \phi(u) du \quad \text{for } x > 0, \\ &= -\frac{1}{x} \int_x^\infty \phi'(u) du \quad \text{by (4),} \\ &= \frac{1}{x} \phi(x). \end{aligned}$$

For the left-hand inequality, we consider

$$\begin{aligned}
 (6) \quad 1 - \Phi(x) &= \int_x^\infty \phi(u) du = - \int_x^\infty \frac{\phi'(u)}{u} du \quad \text{by (4)} \\
 &= - \left[\frac{\phi(u)}{u} \right]_x^\infty - \int_x^\infty \frac{\phi(u)}{u^2} du \quad \text{on integrating by parts,} \\
 &= \frac{\phi(x)}{x} + \int_x^\infty \frac{\phi'(u)}{u^3} du \quad \text{by (4),} \\
 &= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} + \int_x^\infty \frac{3\phi(u)}{u^4} du \quad \text{on integrating by parts,} \\
 &\geq \phi(x) \left\{ \frac{1}{x} - \frac{1}{x^3} \right\}.
 \end{aligned}$$

Remark For large x , these bounds are clearly tight.

- (7) **Exercise** The polynomials $H_n(x)$ are known as Hermite (or Chebyshev–Hermite) polynomials. Show that they are orthogonal with respect to $\phi(x)$ over \mathbb{R} , which is to say that

$$\int_{-\infty}^{\infty} H_n(x) H_m(x) \phi(x) dx = \begin{cases} 0 & m \neq n \\ n! & m = n. \end{cases}$$

- (8) **Exercise** Show that the exponential generating function of the H_n is

$$\sum_{n=0}^{\infty} H_n(x) \frac{t^n}{n!} = e^{tx - \frac{1}{2}t^2}.$$

- (9) **Exercise** Show that for $x > 0$

$$\frac{1}{x} - \frac{1}{x^3} < r(x) < \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5}.$$

- (10) **Exercise** Let X have the Weibull distribution $F(x) = 1 - \exp(-(\lambda t)^2)$. Show that

$$\frac{1}{2}\lambda^{-2}t^{-1} - \frac{1}{4}\lambda^{-4}t^{-3} < \mathbf{E}(X - t | X > t) < \frac{1}{2}\lambda^{-2}t^{-1}.$$

7.13 Example: Bertrand's Paradox

- (a) A point P is chosen at random inside a circular disc of radius a . What is the probability that its distance from O , the centre of the disc, is less than d ? Let X be the length of the chord of the disc of which P is the midpoint. Show that $\mathbf{P}(X > \sqrt{3}a) = \frac{1}{4}$.
- (b) Now choose another chord as follows. A point Q is fixed on the circumference of the disc and a point P is chosen at random on the circumference. Let the length of PQ be Y . Show that $\mathbf{P}(Y > \sqrt{3}a) = \frac{1}{3}$.

Solution (a) If P is less than d from the centre, then it lies inside the disc of radius d with area πd^2 . Therefore, the required probability is

$$(1) \quad \pi d^2 / (\pi a^2) = d^2 / a^2.$$

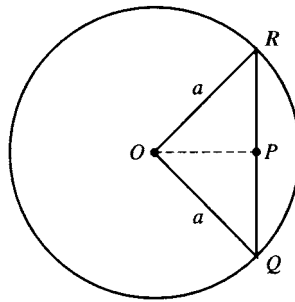


Figure 7.9 Bertrand's paradox. In this case, $X < \sqrt{3}a$ because $OP > \frac{1}{2}a$.

Now $X > \sqrt{3}a$ if and only if the chord RQ subtends an angle greater than $\frac{2\pi}{3}$ at the centre of the disc. This occurs (see Figure 7.9) if and only if OP has length less than $\frac{1}{2}a$.

Hence, by (1),

$$\mathbf{P}(X > \sqrt{3}a) = \frac{(a/2)^2}{a^2} = \frac{1}{4}.$$

(b) As in (a), we observe that $Y > \sqrt{3}a$ if and only if PQ subtends an angle greater than $2\pi/3$ at O . This occurs if and only if P lies on the dashed interval of the circumference of the disc in Figure 7.10. Because this interval is one-third of the circumference, $\mathbf{P}(Y > \sqrt{3}a) = \frac{1}{3}$.

- (2) **Exercise** In part (b), suppose that Q is picked at random as well as P . What is $\mathbf{P}(Y > \sqrt{3}a)$?
- (3) **Exercise** A point P is picked at random on an arbitrarily chosen radius of the disc. Let Z be the length of the chord of which P is the midpoint. Show that $\mathbf{P}(Z > \sqrt{3}a) = \frac{1}{2}$.
- (4) **Exercise** A point Q is fixed on the circumference. The chord is drawn, which makes an angle Θ with the tangent at Q , where Θ is uniform on $(0, \pi)$. If the length of this chord is W , show that $\mathbf{P}(W > \sqrt{3}a) = \frac{1}{3}$.
Is it just a coincidence that this answer is the same as (b) above?

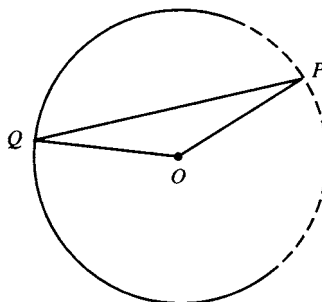


Figure 7.10 Bertrand's paradox. In this case, $Y > \sqrt{3}a$.

7.14 Example: Stock Control

A manufacturer of bits and bobs has a shop. Each week it is necessary to decide how many bits to deliver to the shop on Monday, in light of the following information.

- (i) Delivering y bits costs c per bit, plus a fixed delivery charge k .
- (ii) Any bit unsold at the weekend has to be packed, stored, insured, and discounted over the weekend, at a total cost of h per bit.
- (iii) If the shop sells every bit before the weekend, then further customers that week are supplied by post at the end of the week; this costs p per bit, due to postage, packing, paperwork, and other penalties, and $p > c$.
- (iv) The demand Z for bits each week is a random variable with density $f(z)$ and distribution $F(z)$ where $F(0) = 0$.

If the manager seeks to minimize the expected costs of her decision and she has x bits in the shop over the weekend, approximately how many bits should she order on Monday morning?

Note that the customer pays the same whether the bit comes from the shop or factory. Note also that the problem implicitly assumes that we are content with a continuous approximation to what is actually a discrete problem.

Solution If nothing is delivered, then costs are

$$\begin{aligned} p(Z - x) & \text{ if } Z > x; \\ h(x - Z) & \text{ if } Z < x. \end{aligned}$$

Hence, expected costs are

$$(1) \quad \lambda(x) = p \int_x^\infty (z - x)f(z)dz + h \int_0^x (x - z)f(z)dz.$$

If $y - x$ bits are delivered, to bring the stock of bits to y , then expected costs are

$$(2) \quad \mu(x, y) = k + c(y - x) + \lambda(y).$$

Now

$$\frac{\partial \mu}{\partial y} = c + \lambda'(y) = c + hF(y) - p(1 - F(y))$$

and

$$\frac{\partial^2 \mu}{\partial y^2} = (h + p)f(y) \geq 0.$$

Because $\mu'(0) < 0$, and $\mu'(y) > 0$ for large y , it follows that $\mu(x, y)$ has a unique minimum at the value \hat{y} such that

$$(3) \quad F(\hat{y}) = \frac{p - c}{p + h}.$$

Thus, if any delivery is made, the expected total costs are minimized by choosing $y = \hat{y}$, and the minimum is

$$\mu(x, \hat{y}) = k + c(\hat{y} - x) + \lambda(\hat{y}).$$

The only alternative is to have no delivery, with expected total cost $\lambda(x)$. Hence, the optimal policy is to have no delivery when $x > \hat{y}$ or

$$\lambda(x) \leq k + c\hat{y} + \lambda(\hat{y}) - cx,$$

and to deliver $\hat{y} - x$ when $x < \hat{y}$ and

$$\lambda(x) > k + c\hat{y} + \lambda(\hat{y}) - cx.$$

Now, if we set $g(x) = \lambda(x) + cx$, we have

$$g'(x) = c - p + (h + p)F(x)$$

and

$$g''(x) = (h + p)F(x) \geq 0.$$

Because $g'(0) < 0$ and $g'(\hat{y}) = 0$, it follows that there is a unique point \hat{x} such that

$$(4) \quad g(\hat{x}) = \lambda(\hat{x}) + c\hat{x} = k + c\hat{y} + \lambda(\hat{y}).$$

Hence, the optimal policy takes the simple form:

Deliver no bits if $x \geq \hat{x}$, or

Deliver $\hat{y} - x$ bits if $x < \hat{x}$,

where \hat{y} satisfies (3) and \hat{x} satisfies (4).

(5) **Exercise** What is the optimal policy if the fixed delivery cost k is zero?

(6) **Exercise** Suppose that the postal deliveries also have a setup cost, so that posting y bits costs $m + py$. If demand is exponentially distributed with distribution

$$F(x) = \begin{cases} 1 - e^{-\lambda(x-a)}; & x \geq a \\ 0; & x < a, \end{cases}$$

find the optimal delivery policy.

7.15 Example: Obtaining Your Visa

A certain consular clerk will answer the telephone only on weekdays at about 10.00 A.M. On any such morning, it is an even chance whether he is at his desk or not; if he is absent no one answers, and days are independent. The line is never engaged.

If he is at his desk, the time T that he takes to answer the telephone is a random variable such that

$$(1) \quad \mathbf{P}(T \leq t) = \begin{cases} 0 & t \leq 1 \\ 1 - t^{-1} & t > 1. \end{cases}$$

- (a) If you telephone this clerk one morning, and do not hang up, what is the probability that the telephone rings for at least a time s ?
- (b) You adopt the following procedure. Each day until you are successful you telephone the clerk and hang up at time s if he has not answered by then.

Show that to minimize the expected time you spend listening to the ringing tone, you should choose s to be the unique positive root s_0 of $\log s = (s + 1)(s - 2)$.

Solution (a) Let R be the ringing time. Conditioning on whether the clerk is there or not, we have

$$(2) \quad \begin{aligned} \mathbf{P}(R > s) &= \frac{1}{2}\mathbf{P}(R > s \mid \text{absent}) + \frac{1}{2}\mathbf{P}(R > s \mid \text{present}) \\ &= \begin{cases} 1 & \text{for } s < 1 \\ \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{s} & \text{for } s \geq 1, \text{ by (1)}. \end{cases} \end{aligned}$$

(b) If your call is successful, then the expected time for which the telephone rings is

$$(3) \quad \begin{aligned} \mathbf{E}(R \mid R < s) &= \int_0^s \mathbf{P}(R > x \mid R < s) dx \quad \text{by (7.4.12)} \\ &= \int_0^s \frac{\mathbf{P}(x < R < s) dx}{\mathbf{P}(R < s)} = 1 + \int_1^s \frac{s}{s-1} (x^{-1} - s^{-1}) dx \\ &= \frac{s \log s}{s-1}, s > 1. \end{aligned}$$

The number of unsuccessful calls has a geometric mass function with parameter $\rho = \frac{1}{2}(1 - \frac{1}{s})$, and expectation

$$(4) \quad \frac{1 - \rho}{\rho} = \frac{\frac{1}{2}(1 + \frac{1}{s})}{\frac{1}{2}(1 - \frac{1}{s})} = \frac{s + 1}{s - 1}.$$

Hence, the expected time spent listening to the ringing tone is

$$(5) \quad \omega(s) = \frac{s(s + 1)}{s - 1} + \frac{s \log s}{s - 1}.$$

Differentiating with respect to s gives $\omega'(s) = (s - 1)^{-2}(s^2 - s - 2 - \log s)$. Thus, a stationary value in $(1, \infty)$ occurs at a zero of $s^2 - s - 2 - \log s$. That is where $(s - 2)(s + 1) = \log s$.

- (6) **Exercise** Show that there is just one such zero, and by inspection of (5) this stationary value is a minimum ω_{\min} .
- (7) **Exercise** Show that $\omega(s) \leq 2s^2/(s - 1)$ and deduce that $\omega_{\min} \leq 8$.
- (8) **Exercise** More generally, suppose that the clerk is in his office with probability p and that $\mathbf{P}(T \leq x) = F(x)$.

Show that

$$\mathbf{E}(R) = \frac{s}{pF(s)} - \int_0^s \frac{F(x)}{F(s)} dx.$$

When $F(x) = \frac{x}{1+x}$, show that

$$\mathbf{E}(R) = (1+s)((1-p)p^{-1} + s^{-1} \log(1+s)).$$

7.16 Example: Pirates

Expensive patented (or trade marked) manufactures are often copied and the copies sold as genuine. You are replacing part of your car; with probability p you buy a pirate part, with probability $1 - p$ a genuine part. In each case, lifetimes are exponential, pirate parts with parameter μ , genuine parts with parameter λ , where $\lambda < \mu$. The life of the part you install is T . Is T IFR or DFR? Does it make any difference if $\lambda > \mu$? (See Section 7.7 for expansions of the acronyms.)

Solution By conditional probability,

$$\mathbf{P}(T > t) = \bar{F}(t) = pe^{-\mu t} + (1-p)e^{-\lambda t}.$$

Hence, setting $q = 1 - p$, we have

$$(1) \quad r(t) = f(t)/\bar{F}(t) = \frac{\mu p + \lambda q e^{(\mu-\lambda)t}}{p + q e^{(\mu-\lambda)t}} = \lambda + \frac{p(\mu - \lambda)}{p + q e^{(\mu-\lambda)t}}.$$

This decreases as t increases. Hence, your part has DFR.

It makes no difference if $\lambda > \mu$. This is obvious anyway by symmetry, but also $r(t)$ given by (1) decreases as t increases if $\lambda > \mu$.

- (2) **Exercise** What happens if $\lambda = \mu$?
- (3) **Exercise** Suppose the part has survived for a time t after you install it.
 - (a) Show that the probability π that it is a pirate part is given by

$$\pi(t) = \frac{p}{p + (1-p)e^{(\mu-\lambda)t}}.$$

- (b) Find the limit of $\pi(t)$ as $t \rightarrow \infty$, and explain why the answer depends on whether $\lambda > \mu$ or $\lambda < \mu$.
- (4) **Exercise** Let X have density f and m.g.f. $M_X(\theta) = \mathbf{E}(e^{\theta X})$. Show that

$$\frac{d^2}{d\theta^2} \log(M_X(\theta)) > 0.$$

[You have shown that $M_X(\theta)$ is log-convex, if you are interested.]

- (5) **Exercise** Due to variations in the manufacturing process, the lifetime T is exponential with parameter Λ where Λ has density $f(\lambda)$. Use the preceding exercise to show that T is DFR.
- (6) **Exercise** Let T_Λ be a family of random variables indexed by a parameter Λ , where Λ is a random variable with density $f(\lambda)$. Let $M(t)$ be the continuous mixture

$$M(t) = \mathbf{P}(T_\Lambda \leq t) = \int_0^\infty F_{T_\lambda}(t) f(\lambda) d\lambda.$$

Show that if $F_{T_\lambda}(t)$ is DFR for all λ , then $M(t)$ is DFR. [Hint: The Cauchy–Schwarz inequality says that $\mathbf{E}(XY) \leq (\mathbf{E}(X^2)\mathbf{E}(Y^2))^{\frac{1}{2}}$.]

7.17 Example: Failure Rates[‡]

Let T have distribution $F(t)$.

(a) Show that T is IFRA if and only if, for all $0 \leq \alpha \leq 1$,

$$(1) \quad (\bar{F}(t))^\alpha \leq \bar{F}(\alpha t).$$

(b) Show also that if T is IFRA, then it is NBU.

Solution (a) By definition, T is IFRA if $H(t)/t = \frac{1}{t} \int_0^t r(v)dv$ is increasing in t . This is the same as saying that, for all $0 \leq \alpha \leq 1$,

$$\frac{1}{\alpha t} \int_0^{\alpha t} r(v)dv \leq \frac{1}{t} \int_0^t r(v)dv.$$

But, by (7.7.4), this is equivalent to

$$(2) \quad \frac{-1}{\alpha} \log \bar{F}(\alpha t) \leq -\log \bar{F}(t).$$

Now (1) follows as required because e^x is a monotone increasing function of x .

(b) Because $H(t)/t$ is increasing in t , for all $0 \leq \alpha \leq 1$, we have $H(\alpha t) \leq \alpha H(t)$, and $H((1-\alpha)t) \leq (1-\alpha)H(t)$. Hence,

$$(3) \quad H(\alpha t) + H(t - \alpha t) \leq H(t).$$

Setting $\alpha t = s$ gives condition (iii) in Definition 7.7.8 for NBU.

- (4) **Exercise** Show that if T is IFR, then it is IFRA and DMRL.
 (5) **Exercise** Show that if T is NBU or DMRL, then it is NBUE.
 (6) **Exercise** Let T have a gamma density with parameters 2 and λ . Find $H(t)$ and $r(t)$. Is T IFR?

7.18 Example: Triangles

A point P is chosen at random along a rod of length l .

- (a) The rod is bent at P to form a right angle, thus forming the two shorter sides of a right-angled triangle. Let Θ be the smallest angle in this triangle. Find $\mathbf{E}(\tan \Theta)$ and $\mathbf{E}(\cot \Theta)$.
 (b) The rod is now cut into two pieces at P . A piece is picked at random and cut in half. What is the probability that the three pieces of the rod can form a triangle of any kind? Show that, conditional on the event that a triangle can be formed, the probability that it has no obtuse angle is $2(\sqrt{2} - 1)$.

[‡] See Section 7.7 for expansions of the acronyms.

Solution Without loss of generality, we can suppose the rod to be the unit interval, so that the length OP is a random variable X uniformly distributed on $[0, 1]$.

(a) Because Θ is the smallest angle

$$\tan \Theta = \begin{cases} \frac{X}{1-X}; & 0 \leq X < \frac{1}{2} \\ \frac{1-X}{X}; & \frac{1}{2} \leq X \leq 1. \end{cases}$$

Hence,

$$(1) \quad \mathbf{E}(\tan \Theta) = \int_0^{\frac{1}{2}} \frac{x}{1-x} dx + \int_{\frac{1}{2}}^1 \frac{1-x}{x} dx = 2 \log 2 - 1 \simeq 0.39.$$

For variety and instruction, we choose a different method of finding $\mathbf{E}(\cot \Theta)$. Let $Y = \cot \Theta$. Then, for $y \geq 1$,

$$\begin{aligned} F(y) = \mathbf{P}(Y \leq y) &= \mathbf{P}\left(\frac{1}{\tan \Theta} \leq y\right) \\ &= \mathbf{P}\left(X \leq \frac{y}{1+y} \cap X \geq \frac{1}{1+y}\right) \\ &= F_X\left(\frac{y}{1+y}\right) - F_X\left(\frac{1}{1+y}\right) \\ &= +\frac{y-1}{y+1}. \end{aligned}$$

Hence, differentiating

$$(2) \quad f_Y(y) = \frac{2}{(y+1)^2}.$$

Thus,

$$\mathbf{E}(\cot \Theta) = \mathbf{E}(Y) = \int_1^\infty \frac{2y}{(y+1)^2} dy = \infty.$$

(b) Suppose (without loss of generality) that the piece cut in half has length $1 - X$. Then, if it exists, the triangle is isosceles with sides $\frac{1}{2}(1 - X)$, $\frac{1}{2}(1 - X)$, X . This is possible if and only if $\frac{1}{2}(1 - X) > \frac{1}{2}X$, which occurs if and only if $X < \frac{1}{2}$, which has probability $\frac{1}{2}$.

There is an obtuse angle (between the two sides of equal length) if and only if

$$\frac{\frac{1}{2}X}{\frac{1}{2}(1-X)} > \frac{1}{\sqrt{2}},$$

which occurs if and only if $X > \sqrt{2} - 1$.

Hence,

$$\begin{aligned} \mathbf{P}(\text{no obtuse angle} \mid \text{the triangle exists}) &= \frac{\mathbf{P}\left(\{X < \sqrt{2} - 1\} \cap \{X < \frac{1}{2}\}\right)}{\mathbf{P}\left(X < \frac{1}{2}\right)} \\ &= \frac{\mathbf{P}\left(X < \sqrt{2} - 1\right)}{\mathbf{P}\left(X < \frac{1}{2}\right)} = 2(\sqrt{2} - 1). \end{aligned}$$

- (3) **Exercise** What is the distribution of the length $X \wedge (1 - X)$ of the shortest side of the triangle?
 (4) **Exercise** The longest side is $X \vee (1 - X)$. Show that

$$\frac{\mathbf{E}\{X \wedge (1 - X)\}}{\mathbf{E}\{X \vee (1 - X)\}} = \frac{1}{3},$$

where $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$.

- (5) **Exercise** Find $\frac{\mathbf{E}(\sin \Theta)}{\mathbf{E}(\cos \Theta)}$.
 (6) **Exercise** Show that the hypotenuse of the triangle has density

$$f(y) = \frac{2y}{(2y^2 - 1)^{\frac{1}{2}}}, \quad \frac{1}{\sqrt{2}} \leq y \leq 1.$$

- (7) **Exercise** Let X have density

$$f_X(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}; \quad 0 < x < 1.$$

Show that $\mathbf{E}(\cot \Theta)$ is finite if and only if $a > 1$ and $b > 1$. [See Problem 7.3 for the definition of $B(a, b)$.]

7.19 Example: Stirling's Formula

Let $\Gamma(x)$ be the *gamma function* defined by

$$(1) \quad \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

(a) Show that

$$(2) \quad \Gamma(x) = x^{x-\frac{1}{2}} e^{-x} \int_{-x}^{\infty} (1 + ux^{-\frac{1}{2}})^{x-1} e^{-ux^{\frac{1}{2}}} du.$$

(b) Show that for fixed u the integrand converges to $\exp(-\frac{1}{2}u^2)$, as $x \rightarrow \infty$, and deduce that, as $x \rightarrow \infty$,

$$(3) \quad \Gamma(x) e^x x^{-x+\frac{1}{2}} \rightarrow \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}u^2\right) du.$$

You may assume that $\log(1+x) = x - \frac{1}{2}x^2 + O(x^3)$ for $|x| < 1$.

Proof (a) Making the substitution $t = x + ux^{\frac{1}{2}}$ in (1) gives (2).

(b) Let the integrand in (2) be $f(x, u)$. Then for $u < x^{\frac{1}{2}}$,

$$(4) \quad \begin{aligned} \log f(x, u) &= -ux^{\frac{1}{2}} + (x-1)\log(1+ux^{\frac{1}{2}}) = -\frac{1}{2}u^2 - ux^{-\frac{1}{2}} + O(x^{-1}) \\ &\rightarrow -\frac{1}{2}u^2 \quad \text{as } x \rightarrow \infty. \end{aligned}$$

Now, if we were justified in saying that

$$(5) \quad \lim_{x \rightarrow \infty} \int f(x, u) du = \int \lim_{x \rightarrow \infty} f(x, u) du,$$

then (3) would follow from (2), (4), and (5). However, it is a basic result in calculus[†] that if $0 \leq f(x, u) \leq g(u)$, where $\int_{-\infty}^{\infty} g(u) du < \infty$, then (5) is justified. All we need to do is find a suitable $g(u)$.

First, for $x^{\frac{1}{2}} > 1$ and $u \geq 0$,

$$e^u f(x, u) = e^{-u(x^{1/2}-1)}(1+ux^{\frac{1}{2}})^{x-1} \rightarrow 0 \quad \text{as } u \rightarrow \infty.$$

Hence, $f(x, u) < M_1 e^{-u}$, $u \leq 0$, for some constant M_1 . Second, for $u < 0$, the function $e^{-u} f(x, u)$ has a maximum where

$$-(1+x^{\frac{1}{2}}) + (x-1)(u+x^{\frac{1}{2}})^{-1} = 0,$$

that is, at $u = -1$. Hence,

$$\begin{aligned} f(x, u) &\leq e^u \max_{u < 0} \{e^{-u} f(x, u)\} = e^u (1-x^{-\frac{1}{2}})^{x-1} \exp(1+x^{\frac{1}{2}}) \\ &\rightarrow e^u e^{\frac{1}{2}} \quad \text{as } x \rightarrow \infty. \end{aligned}$$

Hence, for some constant M_2 , $f(x, u) < M_2 e^u$, $u < 0$. Therefore, $f(x, u) < M e^{-|u|}$ for some M , and we have our $g(u)$.

(6) **Exercise** Show that $\lim_{n \rightarrow \infty} n! e^n n^{-n-\frac{1}{2}} = (2\pi)^{\frac{1}{2}}$.

(7) **Exercise** Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = 0$. Given that $S_{2n} = 0$, find the probability P_b that $S_r = b$ for some r such that $0 \leq r \leq 2n$. Show that if $n \rightarrow \infty$ and $b \rightarrow \infty$ in such a way that $b = yn^{\frac{1}{2}}$, then $P_b \rightarrow e^{-y^2}$.

(8) **Exercise: de Moivre–Laplace Theorem** Let S_n be binomial with parameters n and p . Define

$$Y_n = \frac{S_n - np}{(npq)^{\frac{1}{2}}}, \quad q = 1 - p, \quad \text{and} \quad y_k = \frac{k - np}{(npq)^{\frac{1}{2}}}.$$

Show that as $n \rightarrow \infty$

$$\mathbf{P}(S_n = k) = \left(\frac{n}{2\pi k(n-k)} \right)^{\frac{1}{2}} \left(\frac{np}{k} \right)^k \left(\frac{nq}{n-k} \right)^{n-k} (1 + o(1)).$$

Deduce that as $n \rightarrow \infty$

$$\mathbf{P}(Y_n = y_k) = \left(\frac{1}{(2\pi npq)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}y_k^2\right) \right) (1 + O(n^{-\frac{1}{2}}))$$

[†]The Dominated Convergence Theorem.

Conclude that for fixed finite a and b , as $n \rightarrow \infty$,

$$\mathbf{P}(a < Y_n < b) \rightarrow \frac{1}{(2\pi)^{\frac{1}{2}}} \int_a^b e^{-\frac{1}{2}y^2} dy = \Phi(b) - \Phi(a).$$

- (9) **Exercise** Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = 0$. Given that $S_{2n} = 2j$, show that the probability that the last visit to the origin was at the $2r$ th step is

$$f_r = \frac{j}{n-r} \binom{n+j}{r} \binom{n-j}{r} / \binom{2n}{2r}, \quad 0 \leq x < n.$$

Show that if r , j , and n all increase in such a way that $\frac{r}{n} \rightarrow x$, $j/\sqrt{n} \rightarrow y$, then

$$\lim_{r \rightarrow \infty} f_r = \frac{y}{\sqrt{\pi x}} \frac{1}{(1-x)^{\frac{3}{2}}} \exp\left(\frac{-xy^2}{1-x}\right), \quad 0 < x < 1.$$

Remark The result of Exercise 6, which is known as Stirling's formula, was established by de Moivre in 1730. The formula actually proved by Stirling in 1730 was

$$(10) \quad n! \left(n + \frac{1}{2}\right)^{-(n+\frac{1}{2})} e^{n+\frac{1}{2}} \rightarrow (2\pi)^{\frac{1}{2}}, \quad \text{as } n \rightarrow \infty.$$

- (11) **Exercise** Prove (10).

PROBLEMS

- 1 Let $f(x) = c(\alpha, \beta)(x - \alpha)(\beta - x)$. For what values of x and $c(\alpha, \beta)$ can f be a density function?
- 2 Let X have distribution $F(x)$. Show that $\mathbf{P}(X = x) > 0$ if and only if $F(x)$ is discontinuous at x .
- 3 The beta function $B(a, b)$ is given by $B(a, b) = \int_0^1 v^{a-1}(1-v)^{b-1} dv$; $a > 0$, $b > 0$. The beta distribution has density

$$f(x) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1} \quad \text{for } 0 < x < 1.$$

If X has the beta distribution, show that $\mathbf{E}(X) = B(a+1, b)/B(a, b)$. What is $\text{var}(X)$?

- 4 For what value of c is $f = c(\sin x)^\alpha(\cos x)^\beta$; $0 < x < \pi/2$, a density function?
- 5 What is the distribution function of the random variable having the beta density with $a = b = \frac{1}{2}$?
- 6 Let X have the density $f = \exp(-x - \exp(-x))$ for $x \in \mathbb{R}$. What is the distribution function of X ?
- 7 Let X be exponentially distributed with parameter λ . What is the density of $Y = e^{aX}$? For what values of λ does $\mathbf{E}(Y)$ exist?
- 8 Let X have the gamma density with parameters α and λ . Show that

$$\mu_k = \alpha(\alpha+1)\dots(\alpha+k-1)\lambda^{-k}, \quad \text{and } \text{var}(X) = \frac{\alpha}{\lambda^2}.$$

- 9 Let X have the standard normal density, and $a > 0$. Show that

$$\mathbf{P}(X > x + ax^{-1} | X > x) \rightarrow e^{-a} \quad \text{as } x \rightarrow \infty.$$

- 10 (a) Let X have the standard normal density. Show that $|X|$ has distribution function $F = 2\Phi(x) - 1$, for $x > 0$.

- (b) Let X have distribution $F(x)$. What is the distribution of $|X|$? What is the density of $|X|$ if it exists?
- 11 An elastic string has modulus of elasticity λ and natural length l_0 . A mass m is attached to one end, the other being fixed. The period of oscillation of the mass when slightly displaced is $2\pi\sqrt{ml_0/\lambda}$. Suppose that the modulus of elasticity is uniformly distributed on $[a, b]$. What is the density and expectation of the period?
- 12 Let X have density $f(x)$. Construct a simple random variable $S_n(X)$ such that given $\epsilon > 0$, $\mathbf{P}(|S_n(X) - X| > \epsilon) < 2^{-n}$. (Assume X is proper.)
- 13 If X is exponentially distributed find the m.g.f. of X , $\mathbf{E}(e^{tX})$.
- 14 A point Q is chosen at random inside an equilateral triangle of unit side. Find the density of the perpendicular distance X to the nearest side of the triangle.
- 15 For what value of c is $\mathbf{E}((X - c)^2)$ least?
- 16 Suppose a machine's lifetime T has hazard rate $\lambda\sqrt{t}$, where $\lambda > 0$. Find $\mathbf{P}(T > t)$.
- 17 Suppose that X has distribution function

$$F(x) = 1 - \exp\left(-\int_0^x g(u)du\right)$$

for some function $g(\cdot)$. Show that this is possible if and only if $g(u) \geq 0$ and $\int_0^\infty g(u)du = \infty$.

- 18 What are the cumulants of the normal density?
- 19 What are the cumulants of the exponential density?
- 20 You have two independent random variables, each uniform on $(0, 1)$. Explain how you would use them to obtain a random variable X with density

$$f(x) = \frac{3}{5} \left(1 + x + \frac{1}{2}x^2\right) \quad \text{for } 0 \leq x \leq 1.$$

- 21 Define $I(a, b) = \int_0^\infty \exp(-a^2u^2 - b^2u^{-2})du$ for $a, b > 0$. Show that:
 (a) $I(a, b) = a^{-1}I(1, ab)$; (b) $\frac{\partial I}{\partial b} = -2I(1, ab)$; (c) $I(a, b) = \frac{\sqrt{\pi}}{2a}e^{-2ab}$.
- 22 Use the result of Problem 21 to find the m.g.f. of the following densities:
 (a) $f(x) = \alpha x^{-\frac{1}{2}} \exp(-\beta/x - \gamma x)$ for $x > 0$; $\beta, \gamma > 0$. What is α ?
 (b) $f(x) = (2\pi x^3)^{-\frac{1}{2}} \exp(-(2x)^{-1})$ for $x > 0$.
- 23 Let X be a standard normal random variable. What is the density of X^{-2} ?
- 24 Let X be a standard normal random variable. Find the m.g.f. and density of X^2 .
- 25 What is the moment generating function of the two-sided exponential density? Where is it defined?
- 26 Let U be uniform on $(0, 1)$. Show that, if $[a]$ denotes the integer part of a , and $0 < p < 1$,

$$X = 1 + \left[\frac{\log U}{\log(1-p)} \right]$$

has a geometric distribution.

- 27 Let U be uniform on $(0, 1)$. Show how to use U to simulate a random variable with density

$$f(x) = \frac{24}{25} \left(1 + \frac{1}{2} \left(x - \frac{1}{2}\right)^2\right); \quad 0 \leq x \leq 1.$$

- 28 Let P and Q be two points chosen independently and uniformly in $(0, a)$. Show that the distance between P and Q has density $2(a-x)a^{-2}$ for $0 < x < a$.
- 29 **Continuous mixture** Let $f(\theta, x)$ be the exponential density with parameter θ for $0 \leq x$. Let $g(\theta) = \nu e^{-\nu\theta}$ for $\theta \geq 0$, $\nu > 0$. Show that $\int_0^\infty f(\theta, x)g(\theta)d\theta$ is a density.
- 30 Let X_n be a Poisson random variable with parameter n . Show that as $n \rightarrow \infty$, $\mathbf{P}(X_n \leq n + \sqrt{nx}) \rightarrow \Phi(x)$.

- 31** Let U be uniform on $(0, 1)$. Show how to simulate a random variable X with the Pareto distribution given by $F(x) = 1 - x^{-d}; x > 1, d > 1$.
- 32** Let X_α have gamma density with parameters α and 1, so $f_\alpha(x) = \frac{x^{\alpha-1}e^{-x}}{(\alpha-1)!}$. Let $\phi_\alpha(x)$ be the density of $Y_\alpha = (X_\alpha - \alpha)\alpha^{-\frac{1}{2}}$. Show that as $\alpha \rightarrow \infty$, $\phi_\alpha(x) \rightarrow \phi(x)$, where $\phi(x)$ is the standard normal density.
- 33** Let X have mean μ and variance σ^2 . Show that $\mathbf{P}(|X - \mu| \leq a\sigma) \geq 1 - a^{-2}$, for any $a > 0$. Let Z be a standard normal random variable, and define $Y = \alpha + \beta Z + \gamma Z^2$. Find the mean and variance of Y , and show that

$$\mathbf{P}\left(|Y - \alpha - \gamma| \leq \frac{1}{2}\alpha\right) \geq 1 - 4\alpha^{-2}(\beta^2 + 2\gamma^2).$$

- 34** Verify the assertions of Theorem 7.4.16.
- 35** Use the indicator function $I(X > x)$ to show that, for a random variable X that is nonnegative,
- (a) $\mathbf{E}X = \int_0^\infty \mathbf{P}(X > x) dx$.
- (b) $\mathbf{E}X^r = \int_0^\infty r x^{r-1} \mathbf{P}(X > x) dx$.
- (c) $\mathbf{E}e^{\theta X} = 1 + \theta \int_0^\infty e^{\theta x} \mathbf{P}(X > x) dx$.
- (d) When $X \geq 0$ is integer valued,

$$\sum_{k=0}^\infty s^k \mathbf{P}(X > k) = \frac{1 - G_X(s)}{1 - s}.$$

- 36** Let X be a standard normal random variable with density $\phi(x)$ and distribution $\Phi(x)$. Show that for $x > 0$,

$$\frac{x}{x^2 + 1} \phi(x) \leq 1 - \Phi(x) \leq \frac{1}{x} \phi(x).$$

[Hint: Consider $f(x) = xe^{-x^2}/2 - (x^2 + 1)[1 - \Phi(x)]$.]

8

Jointly Continuous Random Variables

Some instances of correlation are quite whimsical: thus cats which are entirely white and have blue eyes are generally deaf.

Charles Darwin, Origin of Species

8.1 Joint Density and Distribution

It is often necessary to consider the joint behaviour of several random variables, which may each take an uncountable number of possible values. Just as for discrete random vectors, we need to define a variety of useful functions and develop the appropriate machinery to set them to work. For simplicity in definitions and theorems, we start by considering a pair of random variables (X, Y) taking values in \mathbb{R}^2 . This theoretical outline can be easily extended to larger collections of random variables (X_1, X_2, \dots, X_n) taking values in \mathbb{R}^n , with a correspondingly greater expenditure of notation and space.

As usual, we should start with a sample space Ω , an event space \mathcal{F} , and a probability function \mathbf{P} , such that for all x and y

$$A_{xy} = \{\omega: X \leq x, Y \leq y\} \in \mathcal{F}.$$

Then

$$(1) \quad F(x, y) = \mathbf{P}(A_{xy}) = \mathbf{P}(X \leq x, Y \leq y)$$

is the joint distribution function of X and Y . In fact, we suppress this underlying structure, and begin with random variables X and Y having joint distribution $F(x, y)$ given by (1). A special class of such jointly distributed random variables is of great importance.

(2) **Definition** Let $F(x, y)$ be a joint distribution. Suppose that $\frac{\partial^2 F}{\partial x \partial y}$ exists and is nonnegative, except possibly on a finite collection of lines in \mathbb{R}^2 . Suppose further that the function $f(x, y)$ defined by

$$f(x, y) = \begin{cases} \frac{\partial^2 F}{\partial x \partial y} & \text{where this exists} \\ 0 & \text{elsewhere,} \end{cases}$$

satisfies

$$(3) \quad F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv.$$

Then X and Y , being random variables having the (joint) distribution F , are said to be (jointly) continuous with (joint) density function $f(x, y)$. ▲

The words “joint” and “jointly” are often omitted to save time and trees. Sometimes we write $f_{X,Y}(x, y)$ and $F_{X,Y}(x, y)$ to stress the role of X and Y , or to avoid ambiguity.

- (4) **Example: Uniform Distribution** Suppose you pick a point Q at random in the rectangle $R = (x, y : 0 < x < a, 0 < y < b)$. Then from the properties of the uniform distribution [see (7.8.1)], we have

$$(5) \quad F(x, y) = \begin{cases} 1 & \text{if } x \geq a, y \geq b \\ \frac{xy}{ab} & \text{if } 0 \leq x \leq a, 0 \leq y \leq b \\ \frac{y}{b} & \text{if } x \geq a, 0 \leq y \leq b \\ \frac{x}{a} & \text{if } 0 \leq x \leq a, y \geq b \\ 0 & \text{elsewhere.} \end{cases}$$

Differentiating wherever possible gives

$$(6) \quad \frac{\partial^2 F}{\partial x \partial y} = \begin{cases} \frac{1}{ab} & \text{if } 0 < x < a, 0 < y < b \\ 0 & \text{if } x < 0 \text{ or } x > a \\ & \text{or } y < 0 \text{ or } y > b. \end{cases}$$

Hence, the function

$$(7) \quad f(x, y) = \begin{cases} \frac{1}{ab} & \text{if } 0 < x < a, 0 < y < b \\ 0 & \text{otherwise} \end{cases}$$

satisfies (3), and is the density of X and Y . It is uniformly distributed over the rectangle R . Furthermore, if A is a subset of R with area $|A|$, then using (7.8.1) (and a theorem about double integrals), we have

$$(8) \quad \mathbf{P}((X, Y) \in A) = \frac{|A|}{ab} = \int \int_{(x,y) \in A} f(x, y) dx dy. \quad \bullet$$

In fact, a version of the useful relationship (8) holds true for all densities $f(x, y)$. This is important enough to state formally as a theorem, which we do not prove.

- (9) **Theorem** If X and Y have density $f(x, y)$ and $\mathbf{P}((X, Y) \in A)$ exists, then it is given by the Key Rule:

$$\mathbf{P}((X, Y) \in A) = \iint_{(x,y) \in A} f(x, y) dx dy.$$

Note that the condition that the probability exists is equivalent to saying that $\{\omega: (X(\omega), Y(\omega)) \in A\} \in \mathcal{F}$. This is another demonstration of the fact that, although we can just about suppress $(\Omega, \mathcal{F}, \mathbf{P})$ at this elementary level, further rigorous progress is not possible without bringing the underlying probability space into play. The attractive result (9) may then be proved.

Here is a simple example of Theorem 9 in use.

- (10) **Example** Let X and Y have density

$$f(x, y) = \begin{cases} 8xy & \text{if } 0 < y < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

What are $\mathbf{P}(2X > 1, 2Y < 1)$ and $\mathbf{P}(X + Y > 1)$? Find $F(x, y)$.

Solution Notice that the constraints $2X > 1, 2Y < 1$ require that $(X, Y) \in S$, where S is the square with vertices $(\frac{1}{2}, 0), (1, 0), (\frac{1}{2}, \frac{1}{2}), (1, \frac{1}{2})$. Hence,

$$\mathbf{P}(2X > 1, 2Y < 1) = 8 \int_{\frac{1}{2}}^1 \int_0^{\frac{1}{2}} f(x, y) dy dx = 8 \int_{\frac{1}{2}}^1 x dx \int_0^{\frac{1}{2}} y dy = \frac{3}{8}.$$

Likewise, $X + Y > 1$ if $(X, Y) \in T$, where T is the triangle with vertices $(\frac{1}{2}, \frac{1}{2}), (1, 0), (1, 1)$. Hence,

$$\mathbf{P}(X + Y > 1) = 8 \int_{\frac{1}{2}}^1 \int_{1-x}^x xy dy dx = \frac{5}{6}.$$

Finally,

$$F(x, y) = \int_0^y \int_v^x 8uv du dv = 2x^2y^2 - y^4. \quad \bullet$$

The geometric problems of Section 7.9 can now be reformulated and generalized in this new framework. Obviously, “picking a point Q at random in some region R ” is what we would now describe as picking (X, Y) such that X and Y are jointly uniform in R . More generally, we can allow (X, Y) to have joint density $f(x, y)$ in R .

Example: More Triangles The random variables X and Y have joint density

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

- (a) What is c ?
 (b) What is $F(x, y)$?

- (c) Show that it is possible to construct a triangle with sides X , Y , $2 - X - Y$, with probability one.
- (d) Show that the angle opposite to the side of length Y is obtuse with probability

$$p_0 = c \int_0^1 \frac{x^{a+1} - x^{a+2}}{2-x} dx.$$

- (e) When $a = 0$, show that $p_0 = 3 - 4 \log 2$.

Solution (a) Because $\int \int f(x, y) dx dy = 1$, this entails

$$c^{-1} = \int_0^1 x^a \int_{1-x}^1 dy dx = (a+2)^{-1}.$$

(b) Using (3) gives

$$\begin{aligned} F(x, y) &= \int_{1-y}^x \int_{1-u}^y cu^a dv du = \frac{a+2}{a+1} yx^{a+1} + x^{a+2} - \frac{a+2}{a+1} x^{a+1} \\ &\quad + \frac{1}{a+1} (1-y)^{a+2}. \end{aligned}$$

(c) Three such lengths form a triangle if $X + Y > 2 - X - Y$, $X + 2 - X - Y > Y$, and $Y + 2 - X - Y > X$. But these constraints are just those that define the region in which $f(x, y)$ is nonzero and $\int \int f dx dy = 1$.

(d) if θ is the angle opposite Y , then

$$\cos \theta = \frac{X^2 + (2 - X - Y)^2 - Y^2}{2X(2 - X - Y)} < 0,$$

if θ is an obtuse angle. Hence, in this case,

$$Y > \frac{X^2 - 2X + 2}{2 - X} = g(X), \quad \text{say.}$$

Now $g(x) \geq 1 - x$ (with equality only at $x = 0$). Hence, p_0 is given by

$$\begin{aligned} \mathbf{P}(\theta \text{ is obtuse}) &= \mathbf{P}(Y > g(X)) = \int_0^1 \int_{g(x)}^1 f(x, y) dy dx \\ &= c \int_0^1 (1 - g(x))x^a dx = c \int_0^1 \frac{x - x^2}{2 - x} x^a dx. \end{aligned}$$

(e) When $a = 0$,

$$p_0 = 2 \int_0^1 \frac{x - x^2}{2 - x} dx = 2 \int_0^1 \frac{2x - x^2}{2 - x} + \frac{2 - x}{2 - x} - \frac{2}{2 - x} dx = 3 - 4 \log 2. \quad \bullet$$

Next we record that, as a result of (1) and Theorem 9, $f(x, y)$ and $F(x, y)$ have the following elementary properties, analogous to those of f and F in the discrete case.

First, $F(x, y)$ is obviously nondecreasing in x and y . More strongly, we have

$$(11) \quad \begin{aligned} 0 &\leq \mathbf{P}(a < X \leq b, c \leq Y \leq d) \\ &= \mathbf{P}(a < X \leq b, Y \leq d) - \mathbf{P}(a < X \leq b, Y \leq c) \\ &= F(b, d) - F(a, d) - F(b, c) + F(a, c). \end{aligned}$$

Second, if X and Y are finite with probability 1, then

$$(12) \quad 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u, v) du dv = \lim_{x, y \rightarrow \infty} F(x, y).$$

Third, knowledge of $F(x, y)$ and $f(x, y)$ will also provide us with the separate distributions and densities of X and Y . Thus, the *marginals* are:

$$(13) \quad F_X(x) = \mathbf{P}(X \leq x) = \lim_{y \rightarrow \infty} \mathbf{P}(X \leq x, Y \leq y) = \int_{-\infty}^{\infty} \int_{-\infty}^x f(u, v) dudv,$$

and

$$(14) \quad f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, v) dv.$$

Likewise,

$$(15) \quad f_Y(y) = \int_{-\infty}^{\infty} f(u, y) du$$

and

$$(16) \quad F_Y(y) = \lim_{x \rightarrow \infty} F(x, y).$$

Here are some examples to illustrate these properties. Note that in future we will specify $f(x, y)$ only where it is nonzero.

- (17) **Example** Verify that the function $f(x, y) = 8xy$ for $0 < y < x < 1$ is a density. For what value of c is $f(x, y) = cxy$ for $0 < x < y < 1$, a density? Find the density of X in the second case.

Solution Because $f > 0$ and

$$\int_0^1 \int_0^x 8xy dy dx = \int_0^1 4x^3 dx = 1,$$

f is indeed a density. By symmetry, $c = 8$ in the second case also, and we have

$$f_X(x) = \int_0^1 8xy dy = 4x(1 - x^2). \quad \bullet$$

- (18) **Example** The function $H(x, y) = 1 - e^{-(x+y)}$ for $x > 0, y > 0$, is nondecreasing in x and y , and $0 \leq H \leq 1$. Is it a distribution?

Solution No, because $\frac{\partial^2 H}{\partial x \partial y}$ exists and is negative in $x > 0, y > 0$. Alternatively, note that

$$H(1, 1) - H(1, 0) - H(0, 1) + H(0, 0) = 2e^{-1} - 1 - e^{-2} < 0,$$

which cannot (as it should) be the value of $\mathbf{P}(0 < X \leq 1, 0 < Y \leq 1)$. ●

(19) Example: Bivariate Normal Density Verify that when $\sigma, \tau > 0$,

$$f(x, y) = \frac{1}{2\pi\sigma\tau(1-\rho^2)^{\frac{1}{2}}} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2} \right) \right]$$

is a density for $|\rho| < 1$, and find the marginal densities $f_X(x)$ and $f_Y(y)$.

Solution From (14), if $f(x, y)$ is a density, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \\ &= \frac{1}{2\pi\sigma\tau(1-\rho^2)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2(1-\rho^2)} \left(\left(\frac{y}{\tau} - \frac{\rho x}{\sigma} \right)^2 + \frac{x^2}{\sigma^2} - \frac{\rho^2 x^2}{\sigma^2} \right) \right] dy \end{aligned}$$

Now setting $\frac{y}{\tau} - \frac{\rho x}{\sigma} = u$, and recalling that

$$\int_{-\infty}^{\infty} \exp \left(-\frac{u^2}{2(1-\rho^2)} \right) \tau du = (2\pi(1-\rho^2))^{\frac{1}{2}} \tau$$

yields

$$f_X(x) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp \left(-\frac{x^2}{2\sigma^2} \right).$$

This is the $N(0, \sigma^2)$ density, and so f satisfies (12) and is nonnegative. It is therefore a density. Interchanging the roles of x and y in the above integrals shows that $f_Y(y)$ is the $N(0, \tau^2)$ density. See Example 8.20 for another approach. ●

8.2 Change of Variables

We have interpreted the random vector (X, Y) as a random point Q picked in \mathbb{R}^2 according to some density $f(x, y)$, where (x, y) are the Cartesian coordinates of Q . Of course, the choice of coordinate system is arbitrary; we may for some very good reasons choose to represent Q in another system of coordinates (u, v) , where (x, y) and (u, v) are related by $u = u(x, y)$ and $v = v(x, y)$. What now is the joint density of $U = u(X, Y)$ and $V = v(X, Y)$?

Equally, given a pair of random variables X and Y , our real interest may well lie in some function or functions of X and Y . What is their (joint) distribution?

As we have remarked above, at a symbolic or formal level, the answer is straightforward. For U and V above, and $A = \{x, y: u(x, y) \leq w, v(x, y) \leq z\}$ then, by Theorem 8.1.9,

$$F_{U,V}(w, z) = \int \int_A f_{X,Y}(x, y) dx dy.$$

The problem is to turn this into a more tractable form.

Fortunately, there are well-known results about changing variables within a multiple integral that provide the answer. We state without proof a theorem for a transformation T satisfying the following conditions. Let C and D be subsets of \mathbb{R}^2 . Suppose that T given by

$$T(x, y) = (u(x, y), v(x, y))$$

maps C one–one onto D , with inverse T^{-1} given by

$$T^{-1}(u, v) = (x(u, v), y(u, v)),$$

which maps D one–one onto C . We define the so-called *Jacobian* J as

$$J(u, v) = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u},$$

where the derivatives are required to exist and be continuous in D . Then we have the following result.

- (1) **Theorem** *Let X and Y have density $f(x, y)$, which is zero outside C . Then $U = u(X, Y)$ and $V = v(X, Y)$ have joint density*

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v)) |J(u, v)| \quad \text{for } (u, v) \in D.$$

Here are some examples of this theorem in use.

- (2) **Example** Suppose $Q = (X, Y)$ is uniformly distributed over the circular disc of radius 1. Then X and Y have joint density

$$(3) \quad f(x, y) = \frac{1}{\pi} \quad \text{for } x^2 + y^2 \leq 1.$$

However, it seems more natural to use polar rather than Cartesian coordinates in this case. These are given by $r = (x^2 + y^2)^{\frac{1}{2}}$ and $\theta = \tan^{-1}(y/x)$, with inverse $x = r \cos \theta$ and $y = r \sin \theta$. They map $C = \{x, y: x^2 + y^2 \leq 1\}$ one–one onto $D = \{r, \theta: 0 \leq r \leq 1, 0 < \theta \leq 2\pi\}$.

In this case,

$$J(r, \theta) = \frac{\partial x}{\partial r} \frac{\partial y}{\partial \theta} - \frac{\partial x}{\partial \theta} \frac{\partial y}{\partial r} = r \cos^2 \theta + r \sin^2 \theta = r.$$

Hence, the random variables $R = r(X, Y)$ and $\Theta = \theta(X, Y)$ have joint density given by

$$(4) \quad f_{R,\Theta}(r, \theta) = \frac{r}{\pi} \quad \text{for } 0 \leq r \leq 1, 0 < \theta \leq 2\pi.$$

Notice that $f(r, \theta)$ is not uniform, as was $f(x, y)$. ●

- (5) **Example** Let $Q = (X, Y)$ be uniformly distributed over the ellipse C with boundary satisfying

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,$$

of area $|C|$. What is $\mathbf{P}(X > Y, X > -Y)$?

Solution Here the transformation $x = ar \cos \theta$ and $y = br \sin \theta$ maps the ellipse one-one onto the circular disc with radius 1. Furthermore, $J = abr$. Now X and Y have density

$$f(x, y) = \frac{1}{|C|}, \quad \text{for } (x, y) \in C,$$

so R and Θ have joint density

$$f(r, \theta) = \frac{abr}{|C|}, \quad \text{for } 0 \leq r < 1, 0 < \theta \leq 2\pi.$$

Hence, $|C| = \pi ab$, and

$$\mathbf{P}(X > Y, X > -Y) = \mathbf{P}\left(-1 < \frac{Y}{X} < 1\right) = \mathbf{P}\left(-1 < \frac{b}{a} \tan \Theta < 1\right) = \frac{1}{\pi} \tan^{-1} \frac{a}{b},$$

because Θ is uniform on $(0, 2\pi)$. ●

8.3 Independence

As usual, independence is an extremely important property; its definition is by now familiar.

- (1) **Definition** *Jointly distributed random variables are independent if, for all x and y ,*

$$\mathbf{P}(X \leq x, Y \leq y) = \mathbf{P}(X \leq x)\mathbf{P}(Y \leq y).$$

In terms of distributions, this is equivalent to the statement that

- (2)
$$F(x, y) = F_X(x)F_Y(y).$$
 ▲

For random variables with a density, it follows immediately by differentiating that

- (3)
$$f(x, y) = f_X(x)f_Y(y)$$

if X and Y are independent. Using the basic property of densities (Theorem 8.1.9) now further shows that, if $C = (x, y: x \in A, y \in B)$ and X and Y are independent, then

- (4)
$$\iint_C f(x, y) dx dy = \int_A f_X(x) dx \int_B f_Y(y) dy.$$

(Assuming of course that the integrals exist.)

Finally, if the random variables U and V satisfy $U = g(X)$, $V = h(Y)$, and X and Y are independent, then U and V are independent. To see this, just let

- (5)
$$A = (x: g(x) \leq u) \quad \text{and} \quad B = (y: h(y) \leq v),$$

and the independence follows from (4) and (2). An important and useful converse is the following.

- (6) **Theorem** If X and Y have density $f(x, y)$, and for all x and y it is true that
- $$f(x, y) = f_X(x)f_Y(y),$$
- then X and Y are independent.

The proof follows immediately from a standard theorem on multiple integrals (just consider $\int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$) and we omit it.

- (7) **Example: Uniform Distribution** Let X and Y have the uniform density over the unit circular disc C , namely,

$$f(x, y) = \pi^{-1} \quad \text{for } (x, y) \in C.$$

- (a) Are X and Y independent?
 (b) Find $f_X(x)$ and $f_Y(y)$.
 (c) If $X = R \cos \Theta$, and $Y = R \sin \Theta$, are R and Θ independent?

Solution (a) The set $\{x, y: x \leq -1/\sqrt{2}, y \leq -1/\sqrt{2}\}$ lies outside C , so

$$F\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right) = 0.$$

However, the intersection of the set $\{x: x \leq -1/\sqrt{2}\}$ with C has nonzero area, so

$$F_X\left(-\frac{1}{\sqrt{2}}\right)F_Y\left(-\frac{1}{\sqrt{2}}\right) > 0.$$

Therefore, X and Y are not independent.

(b) By (8.1.14),

$$f_X(x) = \int_{-1}^1 f(x, y) dy = \frac{1}{\pi} \int_{-(1-x^2)^{\frac{1}{2}}}^{(1-x^2)^{\frac{1}{2}}} dy = \frac{2}{\pi}(1-x^2)^{\frac{1}{2}}.$$

Likewise,

$$f_Y(y) = \frac{2}{\pi}(1-y^2)^{\frac{1}{2}}.$$

(c) By Example 8.2.4, R and Θ have joint density

$$f_{R,\Theta}(r, \theta) = \frac{r}{\pi}, \quad \text{for } 0 \leq r < 1, 0 < \theta \leq 2\pi.$$

Hence,

$$f_{\Theta}(\theta) = \int_0^1 f(r, \theta) dr = \frac{1}{2\pi}; \quad 0 < \theta \leq 2\pi,$$

and

$$f_R(r) = \int_0^{2\pi} f(r, \theta) d\theta = 2r; \quad 0 \leq r \leq 1.$$

Hence, $f(r, \theta) = f_\Theta(\theta)f_R(r)$, and so R and Θ are independent. ●

Example: Bertrand's Paradox Again Suppose we choose a random chord of a circle C radius a , as follows. A point P is picked at random (uniformly) inside C . Then a line through P is picked independently of P at random [i.e., its direction Θ is uniform on $(0, 2\pi)$]. Let X be the length of the chord formed by the intersection of this line with the circle. Show that

$$\mathbf{P}(X > a\sqrt{3}) = \frac{1}{3} + \frac{\sqrt{3}}{2\pi}.$$

Solution Let R be the distance from the centre of the circle to P ; by the above, R has distribution given by $\mathbf{P}(R/a \leq r) = r^2; 0 \leq r \leq 1$. Now $X > a\sqrt{3}$, if and only if $2R \sin \Theta < a$, as you can see by inspecting Figure 8.1. Hence,

$$\begin{aligned} \mathbf{P}(X > a\sqrt{3}) &= \frac{2}{\pi} \int_0^{\pi/2} \mathbf{P}\left(R < \frac{a}{2 \sin \theta}\right) d\theta \\ &= \frac{2}{\pi} \int_0^{\pi/6} d\theta + \frac{2}{\pi} \int_{\pi/6}^{\pi/2} \frac{1}{4} \operatorname{cosec}^2 \theta d\theta = \frac{1}{3} + \frac{\sqrt{3}}{2\pi}. \end{aligned}$$

Compare this with the results of Example 7.13. ●

(8) **Example: Normal Densities** Let X and Y be independent with common density $f(x) = k \exp(-\frac{1}{2}x^2)$ for all x .

- Show that $k = (2\pi)^{-\frac{1}{2}}$.
- Show that $X^2 + Y^2$ and $\tan^{-1}(Y/X)$ are independent random variables.
- If $a > 0 < b < c$ and $0 < \alpha < \frac{1}{2}\pi$, find the probability that $b < (X^2 + Y^2)^{\frac{1}{2}} < c$ and $\frac{1}{4}\pi < \tan^{-1}(Y/X) < \frac{1}{2}\pi$, given that $(X^2 + Y^2)^{\frac{1}{2}} < a$, $Y > 0$, and $\tan^{-1}(Y/X) < \frac{1}{3}\pi$.

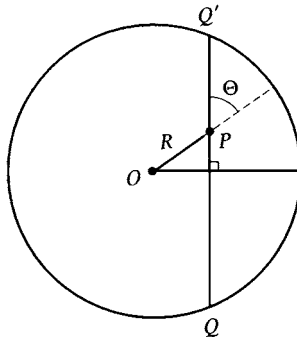


Figure 8.1 Bertrand's paradox.

Solution Because X and Y are independent, they have joint density

$$f(x, y) = k^2 \exp\left(-\frac{1}{2}(x^2 + y^2)\right).$$

Make the change of variables to polar coordinates, so that by Theorem 8.2.1 the random variables $R = (X^2 + Y^2)^{\frac{1}{2}}$ and $\Theta = \tan^{-1}(Y/X)$ have joint density

$$f(r, \theta) = k^2 r \exp\left(-\frac{1}{2}r^2\right) \quad \text{for } 0 \leq r < \infty, 0 < \theta \leq 2\pi.$$

Hence, R has density

$$f_R(r) = r \exp\left(-\frac{1}{2}r^2\right) \quad \text{for } 0 \leq r < \infty,$$

and Θ has density

$$f_\Theta(\theta) = k^2 \quad \text{for } 0 < \theta \leq 2\pi.$$

It follows immediately that

(a) $k^2 = (2\pi)^{-1}$.

(b) $f(r, \theta) = f_R(r)f_\Theta(\theta)$, so that Θ and R are independent by Theorem 6. Hence, Θ and R^2 are independent.

(c) Finally, note that

$$(9) \quad \mathbf{P}\left(\frac{1}{4}\pi < \Theta < \frac{1}{2}\pi, 0 < \Theta < \frac{1}{3}\pi\right) = \mathbf{P}\left(\frac{1}{4}\pi < \Theta < \frac{1}{3}\pi\right) = \frac{\pi}{12}$$

and

$$(10) \quad \mathbf{P}(b < R < c, R < a) = \begin{cases} F_R(c) - F_R(b) & \text{if } c < a \\ F_R(a) - F_R(b) & \text{if } b < a \leq c \\ 0 & \text{otherwise} \end{cases}$$

$$= F_R((a \wedge c) \vee b) - F_R(b),$$

where $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. Now, because R and Θ are independent,

$$\begin{aligned} & \mathbf{P}\left(b < R < c, \frac{1}{4}\pi < \Theta < \frac{1}{2}\pi \mid R < a, 0 < \Theta < \frac{1}{3}\pi\right) \\ &= \frac{\mathbf{P}(b < R < c, R < a)}{\mathbf{P}(R < a)} \frac{\mathbf{P}\left(\frac{1}{4}\pi < \Theta < \frac{1}{2}\pi, 0 < \Theta < \frac{1}{3}\pi\right)}{\mathbf{P}\left(0 < \Theta < \frac{1}{3}\pi\right)} \\ &= \frac{1}{4F_R(a)}(F_R((a \wedge c) \vee b) - F_R(b)), \quad \text{by (9) and (10).} \end{aligned}$$

●

8.4 Sums, Products, and Quotients

We now return to the question of the distribution of functions of random vectors, and take a brief look at some particularly important special cases. Of these, the most important is the sum of two random variables.

Theorem Let X and Y have joint density $f(x, y)$. Show that if $Z = X + Y$, then

$$(1) \quad f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u) du,$$

and that if X and Y are independent, then

$$(2) \quad f_Z(z) = \int_{-\infty}^{\infty} f_X(u) f_Y(z - u) du.$$

Proof First notice that by (8.3.3) the result (2) follows immediately from (1) when X and Y are independent. Turning to the proof of (1), we give two methods of solution.

I Let A be the region in which $u + v \leq z$. Then

$$\begin{aligned} \mathbf{P}(Z \leq z) &= \iint_{(u,v) \in A} f(u, v) du dv = \int_{-\infty}^{\infty} \int_{-\infty}^{z-u} f(u, v) dv du \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^z f(u, w - u) dw du \quad \text{on setting } v = w - u. \end{aligned}$$

Now differentiating with respect to z gives $f_Z(z) = \int_{-\infty}^{\infty} f(u, z - u) du$.

II This time we use the change of variable technique of Section 8.2. Consider the transformation $z = x + y$ and $u = x$, with inverse $x = u$ and $y = z - u$. Here $J = 1$. This satisfies the conditions of Theorem 8.2.1, and so $U = u(X, Y)$ and $Z = z(X, Y)$ have joint density $f(u, z - u)$. We require the marginal density of Z , which is of course just (1). ■

(3) **Example** Let X and Y have the bivariate normal distribution of Example 8.1.19,

$$f(x, y) = \frac{1}{2\pi\sigma\tau(1-\rho^2)^{1/2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2}\right)\right).$$

Find the density of $aX + bY$ for constants a and b .

Remark: Note from 8.3.6 that X and Y are independent if and only if $\rho = 0$.

Solution The joint density of $U = aX$ and $V = bY$ is

$$g(u, v) = \frac{1}{ab} f\left(\frac{u}{a}, \frac{v}{b}\right).$$

Hence, by the above theorem, the density of $Z = U + V = aX + bY$ is

$$(4) \quad f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{ab} f\left(\frac{u}{a}, \frac{z-u}{b}\right) du.$$

Rearranging the exponent in the integrand we have, after a little manipulation,

$$\begin{aligned} & \frac{-1}{2(1-\rho^2)} \left(\frac{u^2}{a^2\sigma^2} - \frac{2\rho u(z-u)}{ab\sigma\tau} + \frac{(z-u)^2}{b^2\tau^2} \right) \\ &= \frac{-1}{2(1-\rho^2)} \left(\alpha \left(u - \frac{\beta}{\alpha} z \right)^2 + \frac{z^2}{\alpha} - \frac{(1-\rho^2)}{a^2b^2\sigma^2\tau^2} \right), \end{aligned}$$

where

$$\alpha = \frac{1}{a^2\sigma^2} + \frac{2\rho}{ab\sigma\tau} + \frac{1}{b^2\tau^2}, \quad \text{and} \quad \beta = \frac{\rho}{ab\sigma\tau} + \frac{1}{b^2\tau^2}.$$

Setting $u = v + \frac{\beta}{\alpha}z$ in the integrand, we evaluate

$$\int_{-\infty}^{\infty} \exp\left(-\frac{\alpha v^2}{2(1-\rho^2)}\right) dv = \left(\frac{2\pi(1-\rho^2)}{\alpha}\right)^{\frac{1}{2}}.$$

Hence, after a little more manipulation, we find that

$$f_Z(z) = \frac{1}{(2\pi\xi^2)^{\frac{1}{2}}} \exp\left(-\frac{z^2}{2\xi^2}\right),$$

where $\xi^2 = a^2\sigma^2 + 2\rho ab\sigma\tau + b^2\tau^2$. That is to say, Z is $N(0, \xi^2)$. ●

One important special case arises when $\rho = 0$, and X and Y are therefore independent. The above result then shows we have proved the following.

(5) **Theorem** *Let X and Y be independent normal random variables having the densities $N(0, \sigma^2)$ and $N(0, \tau^2)$. Then the sum $Z = aX + bY$ has the density $N(0, a^2\sigma^2 + b^2\tau^2)$. (See Example 8.20 for another approach.)* ■

Next we turn to products and quotients.

(6) **Theorem** *Let X and Y have joint density $f(x, y)$. Then the density of $Z = XY$ is*

$$(7) \quad f(z) = \int_{-\infty}^{\infty} \frac{1}{|u|} f\left(u, \frac{z}{u}\right) du,$$

and the density of $W = \frac{X}{Y}$ is

$$f(w) = \int_{-\infty}^{\infty} |u| f(uw, u) du.$$

Proof We use Theorem 8.2.1 again. Consider the transformation $u = x$ and $z = xy$, with inverse $x = u$ and $y = z/u$. Here,

$$J(u, z) = \begin{vmatrix} 1 & 0 \\ \frac{-z}{u^2} & \frac{1}{u} \end{vmatrix} = u^{-1}.$$

This satisfies the conditions of Theorem 8.2.1, and so $U = X$ and $Z = XY$ have joint density

$$(8) \quad f(u, z) = \frac{1}{|u|} f\left(u, \frac{z}{u}\right).$$

The result (7) follows immediately as it is the marginal density of Z obtained from $f(u, z)$.

Alternatively, it is possible to derive the result directly by the usual plod, as follows:

$$\begin{aligned} \mathbf{P}(XY \leq z) &= \mathbf{P}\left(X > 0, Y \leq \frac{z}{X}\right) + \mathbf{P}\left(X < 0, Y \geq \frac{z}{X}\right) \\ &= \int_{-\infty}^0 \int_{z/u}^{\infty} f(u, v) dv du + \int_0^{\infty} \int_{-\infty}^{z/u} f(u, v) dv du \\ &= \int_{-\infty}^0 \int_{-\infty}^z f\left(u, \frac{t}{u}\right) \frac{dt}{(-u)} du + \int_0^{\infty} \int_{-\infty}^z f\left(u, \frac{t}{u}\right) \frac{dt}{u} du \\ &= \int_{-\infty}^z \int_{-\infty}^{\infty} f\left(u, \frac{t}{u}\right) \frac{du}{|u|} dt. \end{aligned}$$

The required density is obtained by comparison of this expression with (7.1.15).

Now we turn to the quotient $W = X/Y$. First, let $V = 1/Y$. Then, by definition,

$$F_{X,V}(x, v) = \mathbf{P}(X \leq x, V \leq v) = \mathbf{P}\left(X \leq x, Y \geq \frac{1}{v}\right) = \int_{-\infty}^x \int_{1/v}^{\infty} f(s, t) ds dt.$$

Hence, on differentiating, the joint density of X and Y^{-1} is given by

$$f_{X,V}(x, v) = \frac{1}{v^2} f\left(x, \frac{1}{v}\right).$$

Now $W = XV$, so by the first part,

$$f_W(w) = \int_{-\infty}^{\infty} \frac{1}{|u|} \frac{u^2}{w^2} f\left(u, \frac{u}{w}\right) du = \int_{-\infty}^{\infty} |v| f(vw, v) dv$$

on setting $u = vw$ in the integrand. Alternatively, of course, you can obtain this by using Theorem 8.2.1 directly via the transformation $w = x/y$ and $u = y$, or you can proceed via the routine plod. ■

As usual, here are some illustrative examples.

(9) **Example** Let X and Y be independent with respective density functions $f_X(x) = xe^{-x/2}$ for $x > 0$ and $f_Y(y) = \pi^{-1}(1 - y^2)^{-1/2}$ for $|y| < 1$.

Show that XY has a normal distribution.

Solution When X and Y are independent, we have $f(x, y) = f_X(x)f_Y(y)$, and Theorem 6 takes the special form

$$\begin{aligned} f(z) &= \int_{-\infty}^{\infty} \frac{1}{|u|} f_X(u) f_Y\left(\frac{z}{u}\right) du = \int_{u>z} \frac{1}{|u|} u e^{-\frac{u^2}{2}} \pi^{-1} \left(1 - \frac{z^2}{u^2}\right)^{-\frac{1}{2}} du \\ &= \frac{1}{\pi} \int_z^{\infty} \frac{e^{-\frac{u^2}{2}}}{(u^2 - z^2)^{\frac{1}{2}}} u du. \end{aligned}$$

Now we make the substitution $u^2 = z^2 + v^2$ to find that

$$f(z) = \frac{1}{\pi} e^{-\frac{z^2}{2}} \int_0^{\infty} e^{-\frac{v^2}{2}} dv = \frac{e^{-\frac{z^2}{2}}}{(2\pi)^{\frac{1}{2}}},$$

which is the $N(0, 1)$ density. ●

- (10) **Example** Let X and Y have density $f(x, y) = e^{-x-y}$ for $x > 0, y > 0$. Show that $U = X/(X + Y)$ has the uniform density on $(0, 1)$.

Solution To use Theorem 6, we need to know the joint density of X and $V = X + Y$. A trivial application of Theorem 8.2.1 shows that X and V have density $f(x, v) = e^{-v}$ for $0 < x < v < \infty$. Hence, by Theorem 6,

$$\begin{aligned} f(u) &= \int_0^{\infty} v e^{-v} dv, & \text{for } 0 < uv < v \\ &= 1, & \text{for } 0 < u < 1. \end{aligned}$$

Alternatively, we may use Theorem 8.2.1 directly by considering the transformation

$$u = \frac{x}{x + y}, \quad v = x + y,$$

with

$$x = uv, \quad y = v(1 - u)$$

and $|J| = v$. Hence, $U = X/(X + Y)$ and $V = X + Y$ have density $f(u, v) = v e^{-v}$, for $v > 0$ and $0 < u < 1$. The marginal density of U is 1, as required. ●

8.5 Expectation

Suppose that the random variable $Z = g(X, Y)$ has density $f(z)$. Then, by definition,

$$\mathbf{E}(Z) = \int_{-\infty}^{\infty} z f(z) dz$$

provided that $\mathbf{E}(|Z|) < \infty$. However, suppose we know only the joint density $f(x, y)$ of X and Y . As we have discovered above, finding the density of $g(X, Y)$ may not be a trivial matter. Fortunately, this task is rendered unnecessary by the following result, which we state without proof.

(1) **Theorem** If X and Y have joint density $f(x, y)$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(u, v)| f(u, v) \, du \, dv < \infty,$$

then

(2)
$$\mathbf{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f(u, v) \, du \, dv.$$

This useful result has the same pleasing consequences as did the corresponding result (Theorem 5.3.1) for discrete random variables.

(3) **Corollary** Let X and Y have finite expectations. Then

- (i) $\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$, for any constants a and b .
- (ii) If $\mathbf{P}(X \leq Y) = 1$, then $\mathbf{E}(X) \leq \mathbf{E}(Y)$.

Suppose further that $\mathbf{E}(X^2)$ and $\mathbf{E}(Y^2)$ are finite. Then

- (iii) $\mathbf{E}(X) \leq \mathbf{E}(|X|) \leq (\mathbf{E}(X^2))^{\frac{1}{2}}$,
- (iv) $\mathbf{E}(XY) \leq (\mathbf{E}(X^2)\mathbf{E}(Y^2))^{\frac{1}{2}}$.

Recall that this last result is the Cauchy–Schwarz inequality.

Finally, suppose that $\mathbf{E}(g(X))$ and $\mathbf{E}(h(Y))$ are finite, and that X and Y are independent. Then

(v) $\mathbf{E}(g(X)h(Y)) = \mathbf{E}(g(X))\mathbf{E}(h(Y))$.

We omit the proofs of these results. Generally speaking, the proofs follow the same line of argument as in the discrete case, with the difference that those proofs used results about rearrangement of sums, whereas these proofs use standard results about multiple integrals.

Definition Some important expectations deserve special mention. Just as we did for discrete random variables, we define the covariance as

$$\text{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))),$$

and the correlation as

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{(\text{var}(X)\text{var}(Y))^{\frac{1}{2}}}.$$

Remark When X and Y are independent, then it follows from Corollary 3(v) that $\text{cov}(X, Y) = \rho(X, Y) = 0$, but not conversely. There is an important exception to this, in that bivariate normal random variables *are* independent if and only if $\rho(X, Y) = 0$. See Examples 8.11 and 8.20 for details

(4) **Example** Let Θ be uniformly distributed over $(0, \alpha)$. Find $\text{cov}(\sin \Theta, \cos \Theta)$, and show that for $\alpha = k\pi$, ($k \neq 0$), $\sin \Theta$ and $\cos \Theta$ are uncorrelated and not independent. ▲

Solution Routine calculations proceed thus:

$$\mathbf{E}(\sin \Theta \cos \Theta) = \frac{1}{\alpha} \int_0^\alpha \frac{1}{2} \sin 2\theta d\theta = \frac{1}{4\alpha}(1 - \cos 2\alpha),$$

and $\mathbf{E}(\sin \Theta) = (1/\alpha)(1 - \cos \alpha)$, and $\mathbf{E}(\cos \Theta) = (1/\alpha) \sin \alpha$. Hence,

$$\text{cov}(\sin \Theta, \cos \Theta) = \frac{1}{4\alpha}(1 - \cos 2\alpha) - \frac{1}{\alpha^2} \sin \alpha(1 - \cos \alpha).$$

This covariance is zero whenever $\alpha = k\pi$, ($k \neq 0$), and so for these values of α , $\rho = 0$. However, $\sin \Theta$ and $\cos \Theta$ are not independent. This is obvious because $\sin \Theta = \cos(\frac{\pi}{2} - \Theta)$, but we can verify it formally by noting that

$$\mathbf{P}\left(\sin \Theta > \frac{3}{4}, \cos \Theta > \frac{3}{4}\right) = 0 \neq \mathbf{P}\left(\sin \Theta > \frac{3}{4}\right) \mathbf{P}\left(\cos \Theta > \frac{3}{4}\right). \quad \bullet$$

Generating functions have been so useful above that it is natural to introduce them again now. Because we are dealing with jointly continuous random variables, the obvious candidate for our attention is a joint moment generating function.

(5) **Definition** Let X and Y be jointly distributed. The joint moment generating function of X and Y is

$$M_{X,Y}(s, t) = \mathbf{E}(e^{sX+tY}). \quad \blacktriangle$$

If this exists in a neighbourhood of the origin, then it has the same attractive properties as the ordinary m.g.f. That is, it determines the joint distribution of X and Y uniquely, and also it does yield the moments, in that

$$(6) \quad \left. \frac{\partial^{m+n}}{\partial s^m \partial t^n} M(s, t) \right|_{s=t=0} = \mathbf{E}(X^m Y^n).$$

Furthermore, just as joint p.g.f.s factorize for independent discrete random variables, it is the case that joint m.g.f.s factorize for independent continuous random variables. That is to say,

$$(7) \quad M_{X,Y}(s, t) = M_X(s)M_Y(t)$$

if and only if X and Y are independent.

We offer no proofs for the above statements, as a proper account would require a wealth of analytical details. Nevertheless, we will use them freely, as required.

(8) **Example** Let X and Y have density $(x, y) = e^{-y}$, $0 < x < y < \infty$. Then

$$M_{X,Y}(s, t) = \int_0^\infty \int_x^\infty e^{sx+ty-y} dy dx = ((1-t)(1-s-t))^{-1}.$$

Hence, differentiating, we obtain

$$\left. \frac{\partial M}{\partial s} \right|_{s=t=0} = 1, \quad \left. \frac{\partial M}{\partial t} \right|_{s=t=0} = 2, \quad \left. \frac{\partial^2 M}{\partial s \partial t} \right|_{s=t=0} = 3.$$

Thus, $\text{cov}(X, Y) = 1$. ●

- (9) **Example** Let X and Y be independent and identically distributed normal random variables. Let $M = \frac{1}{2}(X + Y)$ and $V = (X - M)^2 + (Y - M)^2$. Show that M and V are independent.

Solution We will use an obvious extension of (7). Let $\mathbf{E}(X) = \mu$ and $\text{var}(X) = \sigma^2$. Then consider the joint moment generating function of M , $X - M$, and $Y - M$:

$$\begin{aligned} \mathbf{E}(\exp(sM + t(X - M) + u(Y - M))) \\ &= \mathbf{E}\left(\exp\left(\frac{1}{2}(s + t + u)X + \frac{1}{2}(s - t - u)Y\right)\right) \\ &= \exp\left(\frac{\mu}{2}(s + t + u) + \frac{\sigma^2}{8}(s + t + u)^2\right) \exp\left(\frac{\mu}{2}(s - t - u) + \frac{\sigma^2}{8}(s - t - u)^2\right), \end{aligned}$$

because X and Y are independent,

$$(10) \quad = \exp\left(\mu s + \frac{\sigma^2}{4}s^2\right) \exp\left(\frac{\sigma^2}{4}(t + u)^2\right).$$

Hence, M is independent of the random vector $(X - M, Y - M)$, and so M is independent of V . ●

Remark This remarkable property of the normal distribution extends to any independent collection $(X_i; 1 \leq i \leq n)$ of $N(\mu, \sigma^2)$ random variables, and is known as the independence of sample mean and sample variance property.

- (11) **Example** Let X and Y be independent and identically distributed with mean zero, variance 1, and moment generating function $M(t)$, which is thrice differentiable at 0. Show that if $X + Y$ and $X - Y$ are independent, then X and Y are normally distributed.

Solution By the independence of X and Y ,

$$(12) \quad \begin{aligned} M(s + t)M(s - t) &= \mathbf{E}(e^{(s+t)X})\mathbf{E}(e^{(s-t)Y}) = \mathbf{E}(e^{s(X+Y)+t(X-Y)}) \\ &= \mathbf{E}(e^{s(X+Y)})\mathbf{E}(e^{t(X-Y)}), \end{aligned}$$

by the independence of $X + Y$ and $X - Y$

$$= (M(s))^2 M(t)M(-t),$$

using the independence of X and Y again. Next we note that, by the conditions of the problem, $M'(0) = \mathbf{E}(X) = 0$ and $M''(0) = \mathbf{E}(X^2) = 1$. Now differentiating (12) twice with respect to t , and then setting $t = 0$ gives

$$\begin{aligned} M(s)M''(s) - (M'(s))^2 &= (M(s))^2(M(0)M''(0) - (M'(0))^2) \\ &= (M(s))^2(\mathbf{E}(X^2) - (\mathbf{E}(X))^2) = (M(s))^2. \end{aligned}$$

Integrating this differential equation once gives

$$\frac{M'(s)}{M(s)} = s$$

and integrating again yields $M(s) = \exp(\frac{1}{2}s^2)$. Because this is the moment generating function of the $N(0, 1)$ density, it follows that X and Y have this density. ●

We are of course already aware from Chapter 6 that generating functions are of considerable value in handling sums of independent random variables.

- (13) **Example** Let (X_1, \dots, X_n) be independent having the $N(0, 1)$ density. Show that $Y = \sum_{i=1}^n X_i^2$ has a $\chi^2(n)$ density.

Solution With a view to using moment generating functions, we first find

$$\mathbf{E}(e^{tX_1^2}) = \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}} \exp\left(tx^2 - \frac{1}{2}x^2\right) dx = \frac{1}{(1-2t)^{\frac{1}{2}}}.$$

Hence,

$$(14) \quad \begin{aligned} \mathbf{E}(e^{tY}) &= (\mathbf{E}(e^{tX_1^2}))^n \quad \text{by independence} \\ &= \frac{1}{(1-2t)^{\frac{n}{2}}} \end{aligned}$$

and by (7.1.28) and (7.5.6), this is the m.g.f. of the $\chi^2(n)$ density. Hence, by Theorem 7.5.9, Y has a $\chi^2(n)$ density. ●

Many results about sums of random variables can now be established by methods which, if not trivial, are at least straightforward.

- (15) **Example** Let $(X_k; k \geq 1)$ be independent and identically distributed exponential random variables, with parameter λ . Then for $S_n = \sum_{k=1}^n X_k$,

$$\begin{aligned} \mathbf{E}(e^{tS_n}) &= (\mathbf{E}(e^{tX_1}))^n \quad \text{by independence} \\ &= \left(\frac{\lambda}{\lambda-t}\right)^n \quad \text{by Example 7.5.4,} \end{aligned}$$

and so S_n has a gamma distribution by (7.5.6). ●

8.6 Conditional Density and Expectation

Suppose that X and Y have joint density $f(x, y)$, and we are given the value of Y . By analogy with the conditional mass function that arises when X and Y are discrete, we make the following definition.

- (1) **Definition** If X and Y have joint density $f(x, y)$, then the conditional density of X given $Y = y$ is given by

$$(2) \quad f_{X|Y}(x|y) = \begin{cases} \frac{f(x, y)}{f_Y(y)} & \text{if } 0 < f_Y(y) < \infty \\ 0 & \text{elsewhere.} \end{cases} \quad \blacktriangle$$

We observe immediately that $f_{X|Y}(x|y)$ is indeed a density, because it is nonnegative and

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_Y(y)} dx = \frac{1}{f_Y(y)} \cdot f_Y(y) = 1.$$

The corresponding conditional distribution function is

$$F_{X|Y}(x, y) = \int_{-\infty}^x f_{X|Y}(u|y) du = \mathbf{P}(X \leq x | Y = y),$$

and we have the Key Rule

$$(3) \quad \mathbf{P}(X \in A | Y = y) = \int_{x \in A} f_{X|Y}(x|y) dx.$$

- (4) **Example** Let (X, Y) be the coordinates of the point Q uniformly distributed on a circular disc of unit radius. What is $f_{Y|X}(y|x)$?

Solution Recall that for the marginal density $f_X(x) = (2/\pi)(1 - x^2)^{\frac{1}{2}}$. Hence, by definition,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f(x, y)}{f_X(x)} = \frac{1}{\pi} \frac{\pi}{2(1 - x^2)^{\frac{1}{2}}} \\ &= \frac{1}{2}(1 - x^2)^{-\frac{1}{2}} \quad \text{for } |y| < (1 - x^2)^{\frac{1}{2}}. \end{aligned}$$

This conditional density is uniform on $(-(1 - x^2)^{\frac{1}{2}}, (1 - x^2)^{\frac{1}{2}})$, which is consistent with our earlier observations about conditioning of uniform densities. ●

- (5) **Example** Let X and Y be independent and exponential with parameter λ . Show that the density of X conditional on $X + Y = v$ is uniform on $(0, v)$.

Solution To use (1), we need to take some preliminary steps. First note that the joint density of X and Y is $f(x, y) = \lambda^2 e^{-\lambda(x+y)}$ for $x > 0, y > 0$.

Next we need the joint density of X and $X + Y$ so we consider the transformation $u = x$ and $v = x + y$, with inverse $x = u$ and $y = v - u$, so that $J = 1$. Hence, by Theorem 8.2.1,

$$f_{U, V}(u, v) = \lambda^2 e^{-\lambda v} \quad \text{for } 0 < u < v < \infty.$$

It follows that

$$f_V(v) = \int_0^v \lambda^2 e^{-\lambda v} du = \lambda^2 v e^{-\lambda v},$$

and so by definition

$$f_{U|V}(u|v) = \frac{f(u, v)}{f_V(v)} = \frac{1}{v} \quad \text{for } 0 < u < v.$$

This is the required uniform density. ●

This striking result is related to the lack-of-memory property of the exponential density.

Now, because $f_{X|Y}(x|y)$ is a density it may have an expected value, which naturally enough is called *conditional expectation*.

- (6) **Definition** If $\int_{\mathbb{R}} |x| f_{X|Y}(x|y) dx < \infty$, then the conditional expectation of X given $Y = y$ is given by

$$\mathbf{E}(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x|y) dx. \quad \blacktriangle$$

- (7) **Example (5) Revisited** If X and Y are independent and exponential, then we showed that the density of X given $X + Y = v$ is uniform on $(0, v)$. Hence,

$$\mathbf{E}(X|X + Y = v) = \frac{1}{2}v. \quad \bullet$$

Actually, this is otherwise obvious because, for reasons of symmetry,

$$\mathbf{E}(X|X + Y = v) = \mathbf{E}(Y|X + Y = v),$$

and trivially $\mathbf{E}(X + Y|X + Y = v) = v$. Hence the result follows, provided it is true that for random variables X , Y and V , we have

- (8)
$$\mathbf{E}(X + Y|V = v) = \mathbf{E}(X|V = v) + \mathbf{E}(Y|V = v).$$

In fact, this is true as we now show.

- (9) **Theorem** Let X , Y , and V have joint density $f(x, y, v)$. Then (8) holds.

Proof The joint density of $W = X + Y$ and V is $\int_{-\infty}^{\infty} f(w - u, u, v) du$.

Then, by definition,

$$\mathbf{E}(X + Y|V = v) = \frac{1}{f_V(v)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w f(w - u, u, v) dudw.$$

Now consider the transformation $x = w - u$ and $y = u$, with inverse $w = x + y$ and $u = y$, so that $J = 1$. Changing the variables in the double integral accordingly and using

standard results about such double integrals shows that

$$\begin{aligned} \mathbf{E}(X + Y|V = v) &= \frac{1}{f_V(v)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y)f(x, y, v) dx dy \\ &= \frac{1}{f_V(v)} \int_{-\infty}^{\infty} x f_{X,V}(x, v) dx + \frac{1}{f_V(v)} \int_{-\infty}^{\infty} y f_{Y,V}(y, v) dy \\ &= \mathbf{E}(X|V = v) + \mathbf{E}(Y|V = v). \end{aligned} \quad \blacksquare$$

Next, we make the important observation that by writing

$$(10) \quad \psi(y) = \mathbf{E}(X|Y = y)$$

we emphasize the fact that the conditional expectation of X given Y is a function of Y . If the value of Y is left unspecified, we write

$$\psi(Y) = \mathbf{E}(X|Y)$$

on the understanding that when $Y = y$, $\psi(Y)$ takes the value $\mathbf{E}(X|Y = y)$ defined above.

It is therefore natural to think of $\mathbf{E}(X|Y)$ as a random variable that is a function of Y . (A more rigorous analysis can indeed justify this assumption.) Just as in the discrete case, its expected value is $\mathbf{E}(X)$.

(11) **Theorem** *The expected value of $\psi(Y)$ is $\mathbf{E}(X)$; thus, $\mathbf{E}X = \mathbf{E}(\mathbf{E}(X|Y))$.*

Proof Because $\psi(Y)$ is a function of Y , we can calculate its expected value in the usual way as

$$\begin{aligned} \mathbf{E}(\psi(Y)) &= \int_{-\infty}^{\infty} \psi(y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \quad \text{by Definition 6} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy = \mathbf{E}(X). \end{aligned} \quad \blacksquare$$

We recall an earlier example.

(12) **Example** Let X and Y have density $f(x, y) = 8xy$ for $0 < y < x < 1$. Find $\mathbf{E}(X|Y)$ and $\mathbf{E}(Y|X)$.

Solution Because

$$f_X(x) = \int_0^x 8xy dy = 4x^3,$$

and

$$f_Y(y) = \int_y^1 8xy dx = 4y(1 - y^2),$$

we have $f_{X|Y}(x|y) = 2x/(1 - y^2)$, $y < x < 1$. Hence,

$$\mathbf{E}(X|Y) = \int_Y^1 \frac{2x^2}{(1 - Y^2)^2} dx = \frac{2}{3} \left(\frac{1 - Y^3}{1 - Y^2} \right).$$

Likewise, $f_{Y|X}(y|x) = 2y/x^2$ $0 < y < x$ and, therefore,

$$\mathbf{E}(Y|X) = \int_0^X \frac{2y^2}{X^2} dy = \frac{2}{3} X. \quad \bullet$$

The identity in Theorem 11 can also be used to calculate probabilities by the simple device of letting X be the indicator of the event of interest.

(13) Example Let U and Y have density $f(u, y)$. What is $\mathbf{P}(U < Y)$?

Solution Let X be the indicator of $U < Y$. Then

$$\begin{aligned} \mathbf{P}(U < Y) &= \mathbf{E}(X) = \mathbf{E}(\mathbf{E}(X|Y)) \quad \text{by Theorem 11} \\ &= \mathbf{E} \left(\int_{-\infty}^Y f(u|Y) du \right) = \int_{-\infty}^{\infty} \int_{-\infty}^y f(u, y) du dy \quad \text{by (2)}. \end{aligned}$$

Of course, we could have written this down immediately by Theorem 8.1.9. ●

(14) Example: Bivariate Normal Let X and Y have the bivariate normal density

$$(15) \quad f(x, y) = \frac{1}{2\pi\sigma\tau(1 - \rho^2)^{\frac{1}{2}}} \exp \left(-\frac{1}{2(1 - \rho^2)} \left(\frac{x^2}{\sigma^2} - \frac{2\rho xy}{\sigma\tau} + \frac{y^2}{\tau^2} \right) \right).$$

(a) Find the conditional density of X given $Y = y$.

(b) Find $\mathbf{E}(e^{tXY})$, and hence find the density of $Z = X_1Y_1 + X_2Y_2$, where (X_1, Y_1) is independent of (X_2, Y_2) and each has the density (15).

Solution (a) From Example 8.1.19, we know that Y has the $N(0, \tau^2)$ density. Hence,

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{1}{\sigma(2\pi(1 - \rho^2)^{1/2})} \exp \left(-\frac{1}{2(1 - \rho^2)} \left(\frac{x}{\sigma} - \frac{\rho y}{\tau} \right)^2 \right).$$

Hence, the conditional density of X given $Y = y$ is $N(\rho\sigma y/\tau, \sigma^2(1 - \rho^2))$. Note that if $\rho = 0$, then this does not depend on y , which is to say that X is independent of Y .

(b) By part (a), the conditional moment generating function of X given Y is

$$M_{X|Y}(t) = \exp \left(\frac{\rho\sigma Y}{\tau} t + \frac{1}{2} \sigma^2 (1 - \rho^2) t^2 \right).$$

Hence, by conditional expectation,

$$\begin{aligned} \mathbf{E}(e^{tXY}) &= \mathbf{E}(\mathbf{E}(e^{tXY}|Y)) = \mathbf{E}(M_{X|Y}(tY)) \\ &= \mathbf{E}\left(\exp\left[\left(\frac{\rho\sigma}{\tau}t + \frac{1}{2}\sigma^2(1-\rho^2)t^2\right)Y^2\right]\right) \\ &= \frac{1}{(1-2\rho\sigma\tau t - \sigma^2\tau^2(1-\rho^2)t^2)^{\frac{1}{2}}} \quad \text{using Example 8.5.13,} \end{aligned}$$

and so $X_1Y_1 + X_2Y_2$ has moment generating function,

$$\begin{aligned} M(t) &= (1-2\rho\sigma\tau t - \sigma^2\tau^2(1-\rho^2)t^2)^{-1} \\ &= \frac{1-\rho}{2} \frac{1}{1+\sigma\tau(1-\rho)t} + \frac{1+\rho}{2} \frac{1}{1-\sigma\tau(1+\rho)t}. \end{aligned}$$

Hence, $Z = X_1Y_1 + X_2Y_2$ has an asymmetric bilateral exponential density,

$$(16) \quad f(z) = \begin{cases} \frac{1+\rho}{2} \exp(-\sigma\tau(1+\rho)z) & \text{if } z > 0 \\ \frac{1-\rho}{2} \exp(\sigma\tau(1-\rho)z) & \text{if } z < 0. \end{cases} \bullet$$

We note without proof that $\psi(Y)$ has the useful properties that we recorded in the discrete case. Among the most important is that

$$(17) \quad \mathbf{E}(Xg(Y)|Y) = g(Y)\psi(Y)$$

for any function $g(Y)$ of Y .

Finally, we stress that conditional expectation is important in its own right, it should not be regarded merely as a stage on the way to calculating something else.

For example, suppose that X and Y are random variables, and we want to record the value of X . Unfortunately, X is inaccessible to measurement, so we can only record the value of Y . Can this help us to make a good guess at X ?

First, we have to decide what a “good” guess $g(Y)$ at X is. We decide that $g_1(Y)$ is a better guess than $g_2(Y)$ if

$$(18) \quad \mathbf{E}[(g_1(Y) - X)^2] < \mathbf{E}[(g_2(Y) - X)^2].$$

According to this (somewhat arbitrary) rating, it turns out that the best guess at X given Y is $\psi(Y) = \mathbf{E}(X|Y)$.

(19) Theorem For any function $g(Y)$ of Y ,

$$\mathbf{E}[(X - g(Y))^2] \geq \mathbf{E}[(X - \psi(Y))^2].$$

Proof Using (17), we have

$$(20) \quad \mathbf{E}[(X - \psi)(\psi - g)] = \mathbf{E}[(\psi - g)\mathbf{E}(X - \psi|Y)] = 0.$$

Hence,

$$\begin{aligned} \mathbf{E}[(X - g)^2] &= \mathbf{E}[(X - \psi + \psi - g)^2] = \mathbf{E}[(X - \psi)^2] + \mathbf{E}[(\psi - g)^2] \text{ by (20)} \\ &\geq \mathbf{E}[(X - \psi)^2]. \quad \blacksquare \end{aligned}$$

We conclude this section by recording one more useful property of conditional densities, which may be called the *continuous partition rule*. The proof is an easy *exercise*. For continuous random variables X and Y ,

$$(21) \quad f_X(x) = \int_{\mathbb{R}} f_{X|Y}(x|y) f_Y(y) dy.$$

8.7 Transformations: Order Statistics

We introduced the change of variable technique in Section 8.2. We return to this topic to consider a particularly important class of transformations, namely, those that are linear. Thus, let the random vector (X_1, \dots, X_n) have joint density $f(x_1, \dots, x_n)$. Suppose that for $1 \leq i, j \leq n$, and some constants a_{ij} ,

$$(1) \quad Y_i = \sum_{j=1}^n a_{ij} X_j.$$

What can be said about the joint density of (Y_1, \dots, Y_n) ?

In Section 8.2, we required such transformations to be invertible, and we make the same restriction here. Therefore, we suppose that the matrix $A = (a_{ij})$ has an inverse $A^{-1} = (b_{ij}) = B$. A sufficient condition for this is that the determinant $\det A$ is not zero. Then we have the following useful result, which we state without proof.

(2) **Theorem** *Suppose that (X_1, \dots, X_n) has density $f_X(x_1, \dots, x_n)$, and that (Y_1, \dots, Y_n) is related to (X_1, \dots, X_n) by*

$$Y_i = \sum_{j=1}^n a_{ij} X_j \quad \text{and} \quad X_i = \sum_{j=1}^n b_{ij} Y_j,$$

where $BA = I$, the identity matrix, and $\det A \neq 0$. Then the density of (Y_1, \dots, Y_n) is given by

$$(3) \quad \begin{aligned} f_Y(y_1, \dots, y_n) &= \frac{1}{|\det A|} f_X(x_1(y_1, \dots, y_n) \dots x_n(y_1, \dots, y_n)) \\ &= |\det B| f_X(x_1, \dots, x_n). \end{aligned}$$

(4) **Example: Normal Sample** Let (X_1, \dots, X_n) be independent $N(0, 1)$ random variables, and define

$$(5) \quad Y_j = \sum_{i=1}^n x_j a_{ij} \quad \text{for } 1 \leq j \leq n,$$

where the matrix $A = (a_{ij})$ is an orthogonal rotation with $\det A = 1$ and, denoting the transpose of A by A^T ,

$$AA^T = I = A^T A.$$

- (a) Show that (Y_1, \dots, Y_n) are independent $N(0, 1)$ random variables.
 (b) Deduce that the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

are independent, and that $(n-1)s^2$ has a χ^2 density. ●

Solution (a) It is convenient to use the standard notation for vectors and matrices in problems of this type. Thus, we write x^T for the transpose of x , where

$$(6) \quad x = (x_1, \dots, x_n) = yA^T \quad \text{from (5).}$$

Furthermore,

$$\sum_{i=1}^n x_i^2 = xx^T = yA^T Ay^T = yy^T = \sum_{i=1}^n y_i^2.$$

Hence, by (3), (Y_1, \dots, Y_n) have density

$$(7) \quad f_Y = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2} \sum_{i=1}^n y_i^2\right).$$

Because this factorizes, (Y_1, \dots, Y_n) are independent with the $N(0, 1)$ density.

(b) Now let (a_{ij}) be any rotation such that $a_{1j} = n^{-\frac{1}{2}}$, giving

$$Y_1 = \sum_{j=1}^n \frac{1}{\sqrt{n}} X_j = \sqrt{n} \bar{X}.$$

Then

$$\begin{aligned} (n-1)s^2 &= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \\ &= \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2. \end{aligned}$$

Hence, s^2 is independent of \bar{X} , by the independence of Y_1 and (Y_2, \dots, Y_n) . Finally, because each Y_i is $N(0, 1)$, $(n-1)s^2$ has a $\chi^2(n-1)$ density by Example 8.5.13. See Problem 8.43 for another way to do this. ●

A particularly important linear transformation is the one that places (X_1, \dots, X_n) in non-decreasing order. Thus,

$$\begin{aligned} Y_1 &= \text{smallest of } X_1, \dots, X_n \\ Y_2 &= \text{second smallest of } X_1, \dots, X_n \\ &\vdots \\ Y_n &= \text{largest of } X_1, \dots, X_n. \end{aligned}$$

We assume that each X_k has a density $f(x_k)$, so that the chance of ties is zero. It is customary to use the special notation

$$Y_k = X_{(k)},$$

and then $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are known as the *order statistics* of X_1, \dots, X_n . Now the above transformation is linear, but not one-one. It is in fact many-one; to see this, suppose that $y_1 < y_2 < \dots < y_n$. Then the outcomes

$$X_1 = y_1, X_2 = y_2, \dots, X_n = y_n$$

and

$$X_2 = y_1, X_1 = y_2, \dots, X_n = y_n$$

both yield the same set of order statistics, namely,

$$X_{(1)} = y_1, X_{(2)} = y_2, \dots, X_{(n)} = y_n.$$

However, if $(\pi(1), \dots, \pi(n))$ is any one of the $n!$ distinct permutations of the first n integers and R_π is the region $x_{\pi(1)} < x_{\pi(2)} < \dots < x_{\pi(n)}$, then the transformation

$$x_{(k)} = x_{\pi(k)}; \quad 1 \leq k \leq n$$

is one-one and linear. In the notation of (2), we have

$$a_{ij} = \begin{cases} 1 & \text{if } i = \pi(j) \\ 0 & \text{otherwise,} \end{cases}$$

and $|\det A| = 1$. Therefore, the density of $X_{(1)}, \dots, X_{(n)}$ is $\prod_{i=1}^n f(y_i)$.

Now we observe that X_1, X_2, \dots, X_n lies in just one of the $n!$ regions R_π ; hence, the order statistics have joint density

$$(8) \quad n! \prod_{i=1}^n f(y_i) \quad \text{for } y_1 < y_2 < \dots < y_n.$$

Here are some applications of this useful result.

(9) **Example** Let (X_1, \dots, X_n) be independently and uniformly distributed on $(0, a)$. Then, by (8), their order statistics have the density

$$(10) \quad f = \frac{n!}{a^n} \quad \text{for } y_1 < y_2 < \dots < y_n. \quad \bullet$$

It follows from (8) that we may in principle obtain the marginal density of any subset of the order statistics by performing appropriate integrations. For small subsets, this is actually unnecessary.

(11) **Example** Show that $X_{(k)}$ has density

$$f_{(k)}(y) = k \binom{n}{k} f(y) (1 - F(y))^{n-k} [F(y)]^{k-1}.$$

Solution The event $X_{(k)} \leq y$ occurs if and only if at least k of the X_i lie in $(-\infty, y]$. Hence,

$$(12) \quad F_{(k)}(y) = \sum_{j=k}^n \binom{n}{j} [F(y)]^j (1 - F(y))^{n-j}.$$

Now, differentiating to obtain the density,

$$(13) \quad \begin{aligned} f_{(k)}(y) &= f(y) \sum_{j=k}^n \left[j \binom{n}{j} F^{j-1} (1 - F)^{n-j} - (j+1) \binom{n}{j+1} F^j (1 - F)^{n-(j+1)} \right] \\ &= f(y) k \binom{n}{k} F^{k-1} (1 - F)^{n-k} \end{aligned}$$

by successive cancellation in the sum. ●

8.8 The Poisson Process: Martingales

A recurring idea in previous chapters has been that of a series of events or happenings that may occur repeatedly at random times denoted by T_1, T_2, \dots . For example, we have considered light bulbs that may fail and be replaced at $(T_n; n \geq 1)$, or machine bits that may wear out and be renewed at $(T_n; n \geq 1)$ and so on. Other practical problems may also have this structure; for example, the T_n may be the times at which my telephone rings, cars arrive at the toll booth, meteorites fall from the sky, or you get stung by a wasp.

You can think of many more such examples yourself, and it is clear that it would be desirable to have a general theory of such processes. This is beyond our scope, but we can now consider one exceptionally important special case of such processes.

The basic requirement is that the times between events should be independent and identically distributed random variables ($X_k; k \geq 1$); we assume further that they have an exponential distribution.

(1) **Definition** Let $(X_k; k \geq 1)$ be independent identically distributed exponential random variables with parameter λ . Let $T_0 = 0$, and set

$$T_n = \sum_{k=1}^n X_k; \quad n \geq 1.$$

Define

$$(2) \quad N(t) = \max\{n: T_n \leq t\}; \quad t \geq 0.$$

Then $N(t)$ is a Poisson process with parameter λ . ▲

A couple of remarks are in order here. First, note that $N(t)$ is just the number of happenings or events by time t ; $N(t)$ is constant until an event occurs, when it increases by 1. Second, the collection $(N(t); t \geq 0)$ is an uncountable collection of random variables. We have said nothing about such collections up to now, and so our analysis of $N(t)$ must of necessity be rather informal.

Our first result explains why $N(t)$ is called a Poisson process.

(3) **Theorem** $N(t)$ has mass function

$$f_N(k) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}; \quad k \geq 0.$$

(4) *Proof* First, we note that from Definition 2 of $N(t)$, the event $N(t) \geq k$ occurs if and only if $T_k \leq t$. It follows that

$$(5) \quad \mathbf{P}(N(t) \geq k) = \mathbf{P}(T_k \leq t),$$

and because T_k has a gamma density by (8.5.15) we have

$$\begin{aligned} f_N(k) &= \mathbf{P}(N(t) \geq k) - \mathbf{P}(N(t) \geq k+1) = \mathbf{P}(T_k \leq t) - \mathbf{P}(T_{k+1} \leq t) \\ &= \int_0^t \left(\frac{\lambda^k v^{k-1}}{(k-1)!} - \frac{\lambda^{k+1} v^k}{k!} \right) e^{-\lambda v} dv = \frac{e^{-\lambda t} (\lambda t)^k}{k!} \end{aligned}$$

after an integration by parts. As an alternative, we could argue straight from (5) and (6.1.7) that

$$\begin{aligned} \frac{1 - s\mathbf{E}(s^{N(t)})}{1 - s} &= \sum_0^\infty s^k \mathbf{P}(N(t) \geq k) = 1 + \sum_{k=1}^\infty s^k \int_0^t \frac{\lambda^k v^{k-1}}{(k-1)!} e^{-\lambda v} dv \\ &= 1 + s\lambda \int_0^t e^{\lambda v s - \lambda v} dv = 1 - \frac{s}{1-s} [e^{\lambda v(s-1)}]_0^t \\ &= 1 + \frac{s}{1-s} - \frac{s e^{\lambda t(s-1)}}{1-s} = \frac{1 - s e^{\lambda t(s-1)}}{1-s}, \end{aligned}$$

and the Poisson mass function of $N(t)$ follows. ■

Our next result is one of the most striking and important properties of $N(t)$, from which many other results flow.

(6) **Theorem: Conditional Property of the Poisson Process** *Let $N(t)$ be a Poisson process as defined in Definition 1. Conditional on the event $N(t) = k$, the k random variables*

T_1, \dots, T_k have conditional density

$$(7) \quad f_{T|N=k}(t_1, \dots, t_k) = \frac{k!}{t^k}; \quad 0 < t_1 < t_2 < \dots < t_k \leq t.$$

Before proving (7) let us interpret it. From Example 8.7.9, we recognize that the density $k!/t^k$ is the density of the order statistics of k independent random variables, each uniform on $(0, t)$. Thus, Theorem 6 can be more dramatically expressed: given $N(t) = k$, the k events of the process are independently and uniformly distributed on $(0, t)$.

Proof of (6) Because X_1, \dots, X_k are independent and exponential, they have joint density

$$(8) \quad f(x_1, \dots, x_k) = \lambda^k \exp(-\lambda(x_1 + \dots + x_k)).$$

Next, observe that the transformation

$$t_n = \sum_{i=1}^n x_i; \quad 1 \leq n \leq k+1$$

is linear and invertible with $|J| = 1$. Hence, by Theorem 8.7.2, the random variables $T_n = \sum_{i=1}^n X_i; 1 \leq n \leq k+1$ have joint density

$$(9) \quad f(t_1, \dots, t_{k+1}) = \lambda^{k+1} e^{-\lambda t_{k+1}}; \quad 0 < t_1 < \dots < t_{k+1}.$$

Now

$$\begin{aligned} \mathbf{P}(0 < T_1 < t_1 < T_2 < \dots < T_k < t_k; N(t) = k) \\ &= \mathbf{P}(0 < T_1 < t_1 < \dots < T_k < t_k < t < T_{k+1}) \\ &= \lambda^k t_1(t_2 - t_1) \dots (t_k - t_{k-1}) e^{-\lambda t} \end{aligned}$$

on integrating the density (9). Hence, the conditional distribution of T_1, \dots, T_k given $N(t) = k$ is given by

$$(10) \quad \mathbf{P}(T_1 \leq t_1 < \dots < T_k \leq t_k | N(t) = k) = \mathbf{P}(T_1 \leq t_1 < \dots < t_k; N(t) = k) / \mathbf{P}(N(t) = k) \\ = t_1(t_2 - t_1) \dots (t_k - t_{k-1}) \frac{k!}{t^k}.$$

Now differentiating (10) with respect to all of t_1, \dots, t_k gives (7), as required. ■

As we have remarked, this result finds many applications, see Example 8.17 for some of them. For the moment we content ourselves with showing that $N(t)$ has the so-called *independent increments* property.

(11) Theorem: The Poisson Process has Independent Increments *Let $N(t)$ be a Poisson process, as usual, and let $s < t \leq u < v$. Then $N(t) - N(s)$ is independent of $N(v) - N(u)$.*

Proof Let $W = N(t) - N(s)$ and $Z = N(v) - N(u)$. Then, by conditional expectation,

$$(12) \quad \mathbf{E}(w^W z^Z) = \mathbf{E}[\mathbf{E}(w^W z^Z | N(v))].$$

However, conditional on $N(v) = k$, these events are independently and uniformly distributed in $(0, v)$, whence $(W, Z, k - W - Z)$ has a trinomial distribution with

$$(13) \quad \mathbf{E}(w^W z^Z | N(v) = k) = \left(\left(\frac{t-s}{v} \right) w + \left(\frac{v-u}{v} \right) z + \left(\frac{s+u-t}{v} \right) \right)^k.$$

Hence, combining (12) and (13) and Theorem (3) gives

$$(14) \quad \begin{aligned} \mathbf{E}(w^W z^Z) &= \exp(\lambda w(t-s) + \lambda z(v-u) + \lambda(s+u-t) - \lambda v) \\ &= \exp(\lambda(t-s)(w-1) + \lambda(v-u)(z-1)) = \mathbf{E}(w^W) \mathbf{E}(z^Z), \end{aligned}$$

as required. ■

We may also observe from (14) that because $\mathbf{E}(w^W) = \exp(\lambda(t-s)(w-1))$, it follows that W is Poisson with parameter $\lambda(t-s)$. That is to say, $N(t) - N(s)$ has the same mass function as $N(t-s)$. This property may be called *homogeneity* or the property of *stationary increments*. Note that it is possible, and often convenient, to *define* $N(t)$ as a process with stationary independent Poisson-distributed increments. It is then necessary to prove that interevent times are independent identically distributed exponential random variables.

Now looking back at Section 5.6, we see that the simple random walk also has the property of independent increments. In Section 5.7, we found that martingales were particularly useful in analysing the behaviour of such random walks. We may expect (correctly) that martingales will be equally useful here. First, we need to define what we mean by a martingale for an uncountably infinite family of random variables $(X(t); t \geq 0)$.

(15) **Definition** $(X(t); t \geq 0)$ is a martingale if

(a) $\mathbf{E}|X(t)| < \infty$, for all t .

(b) $\mathbf{E}(X(t)|X(t_1), X(t_2), \dots, X(t_n), X(s)) = X(s)$, for any $0 \leq t_1 < t_2 < \dots < t_n < s < t$.

It is customary and convenient to rewrite (b) more briefly as

$$(b)' \mathbf{E}(X(t)|X(u); u \leq s) = X(s),$$

but note that this is slightly sloppy shorthand; we cannot condition on an infinite number of values of $X(u)$.

A *stopping time* for $X(t)$ is a random variable T taking values in $[0, \infty)$, such that the event $(T \leq t)$ depends only on values of $X(s)$ for $s \leq t$. ▲

There are many technical details that should properly be dealt with here; we merely note that they can all be satisfactorily resolved and ignore them from now on.

Such martingales and their stopping times can behave much like those with discrete parameter. We state this useful theorem without proof, or technical details.

(16) Theorem: Optional Stopping Let $X(t)$ be a continuous parameter martingale and T a stopping time for $X(t)$. Then $X(t \wedge T)$ is a martingale, so

$$\mathbf{E}X(0) = \mathbf{E}X(t \wedge T) = \mathbf{E}X(t).$$

Furthermore, $\mathbf{E}X(T) = \mathbf{E}X(0)$ if any one of the following holds for some constant $K < \infty$:

- (a) T is bounded (i.e., $T \leq K < \infty$).
- (b) $|X(t)| \leq K$ for all t , and $\mathbf{P}(T < \infty) = 1$.
- (c) $\mathbf{E}(\sup X(t \wedge T)) < \infty$, and $\mathbf{P}(T < \infty) = 1$.
- (d) $\mathbf{E}|X(T)| \leq K$, and $\mathbf{P}(T < \infty) = 1$, and $\lim_{t \rightarrow \infty} \mathbf{E}(X(t)I(T > t)) = 0$.

Note that similar appropriate definitions and theorems apply to sub- and supermartingales with continuous parameter, but we do not explore that further here. Note further that, just as in the discrete-time case, a popular technique when $\mathbf{P}(T < \infty) = 1$ is to apply part (a) of the theorem at $T \wedge t$, allow $t \rightarrow \infty$, and use either the Dominated or Monotone Convergence theorem, as appropriate.

(17) Example: Poisson Martingales If $(N(t); t \geq 0)$ is a Poisson process with parameter λ , then the following are all martingales:

- (a) $U(t) = N(t) - \lambda t$.
- (b) $V(t) = (U(t))^2 - \lambda t$.
- (c) $W(t) = \exp\{-\theta N(t) + \lambda t(1 - e^{-\theta})\}$, $\theta \in \mathbb{R}$.

To see (a), we calculate

$$\begin{aligned} \mathbf{E}(U(t)|N(u); 0 \leq u \leq s) &= \mathbf{E}(N(t) - N(s) + N(s)|N(u), 0 \leq u \leq s) \\ &= U(s) + \mathbf{E}[N(t) - N(s)] - \lambda(t - s) \\ &= U(s), \end{aligned}$$

where we used the independent increments property. The proof of (b) proceeds likewise:

$$\begin{aligned} \mathbf{E}(V(t)|N(u), 0 \leq u \leq s) &= \\ &= \mathbf{E}[N(t) - \lambda t - N(s) + \lambda s]^2 | N(u); 0 \leq u \leq s] - \lambda t \\ &= [U(s)]^2 + \lambda(t - s) - \lambda t = V(s), \end{aligned}$$

on using the independence of the increments and $\text{var}[N(t) - N(s)] = \lambda(t - s)$. Part (c) is left as an exercise, but see also Example (8.22). ●

8.9 Two Limit Theorems

Perhaps surprisingly (as we are toward the end of the book), this is an appropriate moment to reconsider our basic ideas about chance. Suppose we are given a number n of similar observations or measurements, denoted by x_1, \dots, x_n . For example, the x_i may be the height of each of n men, or they may be the lifetimes of n light bulbs, or they may be the weight of potatoes yielded by each of n plants. By “similar” in this context, we mean that no measurement has any generic reason to be larger or smaller than the others; the

potatoes are of the same variety and grown in the same circumstances; the light bulbs are of the same make and type, and the men are of the same age and race.

It is convenient to have one number that gives an idea of the size of a typical x_i , and a popular candidate for this number is the average \bar{x} given by

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i.$$

One reason for the popularity of \bar{x} is that it is empirically observed that, as n increases, the sequence \bar{x}_n undergoes smaller and smaller fluctuations, and indeed exhibits behaviour of the kind we call convergent.

A special case of such measurements arises when each x_i takes the value 1 or 0 according to whether some event A occurs. Then \bar{x}_n is the proportion of times that A occurs in n trials, and the fact that \bar{x}_n fluctuates less and less as n increases is sometimes used as a basis to justify the axioms of probability.

Of course, in mathematical terms, we think of x_i as the outcome of some random variable X_i . It follows that, if our theory of probability is as relevant as we have claimed, then the sequence

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

ought also to exhibit a similar kind of regularity in the long run as $n \rightarrow \infty$. What kind might there be?

To gain some insight into the problem, consider the case when each x_i is 1 if A occurs, and 0 otherwise. In the mathematical formulation of this, X_i is the indicator of the event A , and we assume the X_i are independent. Then $\sum_1^n X_i$ is a binomial random variable with parameters n and p , and we have shown in Examples 4.18 and 7.5.11 that, as $n \rightarrow \infty$,

$$(1) \quad \mathbf{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i - p \right| > \epsilon \right) \rightarrow 0$$

and

$$(2) \quad \mathbf{P} \left(\frac{1}{(npq)^{\frac{1}{2}}} \sum_{i=1}^n (X_i - p) \leq x \right) \rightarrow \Phi(x),$$

where $\Phi(x)$ is the standard normal distribution. (Indeed, we proved something even stronger than (1) in Example 4.18 and Theorem 5.8.8.)

It seems that $n^{-1} \sum_1^n (X_i - \mathbf{E}(X_1))$ is settling down around $\mathbf{E}(X_i) = p$ and that the distribution of $(n \operatorname{var}(X_1))^{-\frac{1}{2}} \sum_1^n (X_i - \mathbf{E}(X_1))$ is getting closer to the standard normal distribution $\Phi(x)$. More generally, we showed in Theorem 5.8.6 that (1) holds for any collection of independent discrete random variables with the same mean and variance. We called this the weak law of large numbers.

This is deliberately vague and informal, but it should now seem at least plausible that the following results might be true.

(3) **Theorem** Let $(X_k; k \geq 1)$ be independent and identically distributed random variables with mean μ , variance $\sigma^2 < \infty$, and moment generating function $M_X(t)$, $|t| < a$. Then we have:

(i) **Weak Law of Large Numbers** For $\varepsilon > 0$, as $n \rightarrow \infty$,

$$(4) \quad \mathbf{P} \left(\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mu) \right| > \varepsilon \right) \rightarrow 0.$$

(ii) **Central Limit Theorem** As $n \rightarrow \infty$,

$$(5) \quad \mathbf{P} \left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \leq x \right) \rightarrow \Phi(x) = \int_{-\infty}^x (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} dy.$$

It is a remarkable fact that both of these are indeed true, and we now prove them.

Proof of (4) The essential step here is to recall Chebyshev's inequality, for then we may write, using Theorem 7.4.16,

$$\begin{aligned} \mathbf{P} \left(\frac{1}{n} \left| \sum_1^n (X_i - \mu) \right| > \varepsilon \right) &\leq \varepsilon^{-2} \mathbf{E} \left(n^{-2} \left(\sum_1^n (X_i - \mu) \right)^2 \right) \\ &= \varepsilon^{-2} n^{-2} \mathbf{E} \left(\sum_1^n (X_i - \mu)^2 \right) \quad \text{by independence,} \\ &= \varepsilon^{-2} n^{-1} \sigma^2 \rightarrow 0 \quad \text{as } n \rightarrow \infty \end{aligned}$$

■

Note that the proof here is the same as that of Theorem 5.8.6. So it is not necessary for (4) that the X_i be identically distributed or have an m.g.f., it is sufficient that they have the same mean and variance.

Proof of (5) The essential step here is to recall the continuity theorem (7.5.10), for then we may write

$$\begin{aligned} \mathbf{E} \left(\exp \left[\frac{t}{\sigma\sqrt{n}} \sum_1^n (X_i - \mu) \right] \right) &= \left(\mathbf{E} \left(\exp \left[\frac{t}{\sigma\sqrt{n}} (X_1 - \mu) \right] \right) \right)^n, \quad \text{by independence} \\ &= (M_Y(t/(\sigma\sqrt{n})))^n \quad \text{where } Y = X - \mu, \\ &= \left(1 + \mathbf{E}(Y^2) \frac{t^2}{2\sigma^2 n} + o\left(\frac{t^2}{n}\right) \right)^2 \quad \text{by Theorem 7.5.9} \\ &\rightarrow e^{\frac{1}{2}t^2}, \quad \text{because } \mathbf{E}(Y^2) = \sigma^2. \end{aligned}$$

Now we recall that $e^{\frac{1}{2}t^2}$ is the moment generating function of the standard normal density $\phi(x)$, and (5) follows by the continuity theorem (7.5.10). ■

8.10 Review and Checklist for Chapter 8

In this chapter, we consider the joint behaviour of collections of continuous random variables having joint density functions. We also introduce the joint distribution function, and show how these yield the marginal densities and distributions. The change of variable technique is given and used to study important functions of sets of random variables (including sums, products, quotients, and order statistics). We look at expectation, independence, and conditioning, especially the key concept of conditional expectation. Finally, we discuss the Poisson process and its crucial properties, together with continuous parameter martingales and the optional stopping theorem, and prove simple forms of the weak law of large numbers and the central limit theorem.

We summarize most of these principal properties for the bivariate case (X, Y) . The extension to larger collections of random variables (X_1, X_2, X_3, \dots) ; the multivariate case) is straightforward but typographically tedious.

SYNOPSIS OF FORMULAE:

The random vector (X, Y) is supposed to have joint density $f(x, y)$ and distribution $F(x, y)$.

Key rule:

$$\mathbf{P}[(X, Y) \in B] = \int_{(x,y) \in B} f(x, y) dx dy.$$

Basic rules:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv = \mathbf{P}(X \leq x, Y \leq y).$$

For small h and k ,

$$\mathbf{P}(x < X \leq x + h, y < Y \leq y + k) \simeq f(x, y)hk,$$

and when F is differentiable,

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y).$$

Marginals:

$$f_X(x) = \int_{\mathbb{R}} f(x, y) dy; \quad f_Y(y) = \int_{\mathbb{R}} f(x, y) dx,$$

$$F_X(x) = F(x, \infty); \quad F_Y(y) = F(\infty, y).$$

Functions:

$$\mathbf{P}(g(X, Y) \leq z) = \int \int_{g \leq z} f(x, y) dx dy.$$

In particular,

$$f_{X+Y}(z) = \int_{\mathbb{R}} f(x, z-x) dx.$$

Transformations:

More generally, suppose $(u, v) = (u(x, y), v(x, y))$ defines a one–one invertible function; let

$$J(u, v) = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial y}{\partial u} \frac{\partial x}{\partial v},$$

where the derivatives are continuous in the domain of (u, v) . Then the random variables $(U, V) = (u(X, Y), v(X, Y))$ are jointly continuous with density

$$f_{U,V}(u, v) = f_{X,Y}(x(u, v), y(u, v))|J(u, v)|.$$

Independence: X and Y are independent, if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \quad \text{for all } x, y,$$

or

$$f_{X,Y}(x, y) = f_X(x)f_Y(y), \quad \text{for all } x, y.$$

Expectation: If $\int \int |g|f(x, y)dx dy < \infty$, then $g(X, Y)$ has an expected value given by

$$\mathbf{E}g(X, Y) = \int \int g(x, y)f(x, y)dx dy.$$

Moments: Joint moments, covariance, and correlation are defined as they were in the discrete case.

Joint generating functions: The joint probability generating function of integer-valued X and Y is

$$G_{X,Y}(x, y) = \mathbf{E}(x^X y^Y).$$

The joint moment generating function of X and Y is

$$M_{X,Y}(s, t) = \mathbf{E}(e^{sX+tY}) = G_{X,Y}(e^s, e^t).$$

Moments are obtained by appropriate differentiation, so

$$\text{cov}(X, Y) = \begin{cases} \frac{\partial^2 G}{\partial x \partial y}(1, 1) - \frac{\partial G}{\partial x}(1, 1) \frac{\partial G}{\partial y}(1, 1) & \text{if } X, Y \text{ are discrete} \\ \frac{\partial^2 M}{\partial s \partial t}(0, 0) - \frac{\partial M}{\partial s}(0, 0) \frac{\partial M}{\partial t}(0, 0) & \text{in any case.} \end{cases}$$

Random variables X and Y are independent if and only if the moment generating function factorizes as a product of separate functions of s and t for all s, t ; thus,

$$M(s, t) = M(s, 0)M(0, t).$$

Conditioning: The conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \begin{cases} \frac{f(x, y)}{f_Y(y)}, & 0 < f_Y(y) < \infty \\ 0, & \text{otherwise.} \end{cases}$$

The Key Rule is

$$\mathbf{P}(X \in A|Y = y) = \int_{x \in A} f_{X|Y}(x|y) dx,$$

the continuous partition rule is

$$f_X(x) = \int_{\mathbb{R}} f_{X|Y}(x|y) f_Y(y) dy,$$

and the conditional distribution function is

$$F_{X|Y}(x|y) = \int_{-\infty}^x f_{X|Y}(x|y) dx = \mathbf{P}(X < x|Y = y).$$

If $\int |x| f_{X|Y}(x|y) dx < \infty$, then the conditional expectation of X given $Y = y$ is

$$\mathbf{E}(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x|y) dx.$$

As Y varies, this defines a function $\mathbf{E}(X|Y)$, where $\mathbf{E}(X|Y) = \mathbf{E}(X|Y = y)$ when $Y = y$.

Key theorem for conditional expectation:

$$\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}X.$$

This has the same properties as the conditional expectation in the discrete case.

Limit Theorems: We established the Weak Law of Large Numbers and the Central Limit Theorem in (8.9.3).

Multivariate normal density: Jointly normal random variables are particularly important. Recall that the standard normal density is

$$\phi(x) = f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \quad -\infty < x < \infty.$$

If the random variables X and Y are jointly distributed, and $aX + bY$ has a normal distribution for all choices of the constants a and b , then X and Y are said to have a bivariate normal, or *binormal*, density. In particular, X and Y have the standard binormal density with correlation coefficient (or parameter) ρ , if

$$f(x, y) = (2\pi)^{-1} (1 - \rho^2)^{-1/2} \exp\left\{\frac{-1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2)\right\}.$$

Thus X and Y are independent if and only if uncorrelated.

More generally, $\mathbf{X} = (X_1, \dots, X_n)$ is said to be multivariate normal (or *multinormal*) if $\sum_1^n a_i X_i$ has a normal distribution for all choices of $\mathbf{a} = (a_1, \dots, a_n)$. It follows that the multinormal distribution is determined by the means and covariances of (X_1, \dots, X_n) .

To see this, we simply calculate the joint moment generating function of \mathbf{X} :

$$M_{\mathbf{X}}(\mathbf{t}) = \mathbf{E} \exp \left(\sum_1^n t_r X_r \right).$$

This is easy because $\sum_1^n t_r X_r$ is normal with mean $\sum_i t_i \mathbf{E}X_i$ and variance $\sum_{i,j} \text{cov}(X_i, X_j) t_i t_j$. Hence, by Example 7.5.7,

$$M_{\mathbf{X}}(\mathbf{t}) = \exp \left[\sum_i t_i \mathbf{E}X_i + \frac{1}{2} \sum_{i,j} \text{cov}(X_i, X_j) t_i t_j \right],$$

and this determines the joint distribution by the basic property of m.g.f.s.

In particular, X_1, X_2, X_3 have the *standard trivariate normal* distribution (or trinormal) when $\text{var}X_1 = \text{var}X_2 = \text{var}X_3 = 1$, $\mathbf{E}X_1X_2 = \rho_{12}$, $\mathbf{E}X_1X_3 = \rho_{23}$, $\mathbf{E}X_3X_1 = \rho_{31}$, and

$$f(x_1, x_2, x_3) = (2\pi)^{-3/2} A^{-1/2} \exp \left\{ -\frac{1}{2A} \sum_j \sum_k a_{jk} x_j x_k \right\},$$

where $a_{11} = 1 - \rho_{23}^2$, $a_{22} = 1 - \rho_{31}^2$, $a_{33} = 1 - \rho_{12}^2$, $a_{12} = a_{21} = \rho_{31}\rho_{23} - \rho_{12}$, $a_{13} = a_{31} = \rho_{12}\rho_{23} - \rho_{31}$, $a_{23} = a_{32} = \rho_{12}\rho_{31} - \rho_{23}$, and $A = 1 - \rho_{12}^2 - \rho_{23}^2 - \rho_{31}^2 + 2\rho_{12}\rho_{23}\rho_{31}$. The joint moment generating function of X_1, X_2, X_3 is

$$M(t_1, t_2, t_3) = \exp \left\{ \frac{1}{2} (t_1^2 + t_2^2 + t_3^2 + \rho_{12}t_1t_2 + \rho_{23}t_2t_3 + \rho_{31}t_3t_1) \right\}.$$

Checklist of Terms for Chapter 8

- 8.1 joint distribution function
 - joint density function
 - marginals
 - bivariate normal density
- 8.2 change of variable formula
- 8.3 independence and factorization
- 8.4 sums, products, and quotients
- 8.5 expectation
 - Cauchy–Schwarz inequality
 - independence and expectation
 - covariance and correlation
 - joint moment generating function
 - normal sample
- 8.6 conditional density
 - conditional distribution
 - conditional expectation
- 8.7 normal sample
 - order statistics
- 8.8 Poisson process
 - conditional property

- independent increments
- martingales
- optional stopping theorem
- 8.9 weak law of large numbers
- central limit theorem

WORKED EXAMPLES AND EXERCISES

8.11 Example: Bivariate Normal Density

Let X and Y have the standard bivariate normal joint density

$$f(x, y) = \frac{1}{2\pi(1 - \rho^2)^{\frac{1}{2}}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1 - \rho^2)}\right).$$

Show that the joint moment generating function of X and Y is

$$\exp\left(\frac{1}{2}(s^2 + 2\rho st + t^2)\right).$$

Solution We are asked to find

$$(1) \quad M = \mathbf{E}(e^{sX+tY}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)e^{sx+ty} dx dy.$$

After a little thought, we observe that the terms in the exponents in the integrand can be rearranged to give

$$M = \frac{1}{2\pi(1 - \rho^2)^{\frac{1}{2}}} \int \int \exp\left(-\frac{x^2}{2} + x(s + t\rho) - \frac{1}{2}\left(\frac{y - \rho x}{(1 - \rho^2)^{\frac{1}{2}}}\right)^2 + \frac{y - \rho x}{(1 - \rho^2)^{\frac{1}{2}}}t(1 - \rho^2)^{\frac{1}{2}}\right) dx dy.$$

This suggests that we make the change of variables $u = x, v = (y - \rho x)/(1 - \rho^2)^{\frac{1}{2}}$ in the integral. This map is one–one, and $J = (1 - \rho^2)^{\frac{1}{2}}$. Hence,

$$M = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}u^2 + (s + t\rho)u - \frac{1}{2}v^2 + t(1 - \rho^2)^{\frac{1}{2}}v\right) dudv.$$

Because the integrand factorizes, we now recognize the right-hand side as being equal to

$$\mathbf{E}(e^{(s+t\rho)U})\mathbf{E}(e^{t(1-\rho^2)^{\frac{1}{2}}V}),$$

where U and V are standard normal random variables. But we know the m.g.f. $\mathbf{E}(e^{tV})$ of a standard normal random variable V to be $e^{\frac{1}{2}t^2}$. Hence,

$$(2) \quad M = e^{\frac{1}{2}(s+t\rho)^2} e^{\frac{1}{2}t^2(1-\rho^2)},$$

as required.

- (3) **Exercise** Find the conditional m.g.f. of Y given X .
- (4) **Exercise** Use (3) to find $\mathbf{E}(e^{sX+tY})$. Show that $\rho(X, Y) = \text{cov}(X, Y) = \rho$. Deduce that X and Y are independent if and only if $\text{cov}(X, Y) = 0$.
- (5) **Exercise** Let X_1, X_2, \dots, X_n be independent standard normal variables. Let $W = \sum_{i=1}^n \alpha_i X_i$ and $Z = \sum_{i=1}^n \beta_i X_i$. When are Y and Z independent?
- (6) **Exercise** Find the distribution of $aX + bY$, where X and Y have the bivariate normal distribution.

Remark See Example 8.20 for another approach.

8.12 Example: Partitions

- (a) The random variables X and Y are independently and uniformly distributed on $(0, a)$. Find the density of U, V , and W , where $U = \min\{X, Y\}$, $V = |X - Y|$, and $W = a - \max\{X, Y\}$.
- (b) Use this to show that if three points are picked independently and uniformly on the perimeter of a circle of radius r , then the expected area of the resulting triangle is $3r^2/(2\pi)$.

Solution (a) We give three methods of solution.

I: *Basic Plot* (i) By independence

$$\mathbf{P}(U \leq u) = 1 - \mathbf{P}(X > u; Y > u) = 1 - \left(\frac{a-u}{a}\right)^2.$$

(ii) By the basic property of densities, if we let C be the set $\{x, y: |x - y| \leq v\}$, then

$$\mathbf{P}(V \leq v) = a^{-2} \int_C dx dy = 1 - \left(\frac{a-v}{a}\right)^2.$$

(iii) By independence,

$$\mathbf{P}(W \leq w) = \mathbf{P}(\max\{X, Y\} \geq a - w) = 1 - \left(\frac{a-w}{a}\right)^2.$$

Hence, U, V , and W have the same density:

$$(1) \quad f(z) = \frac{2}{a^2}(a - z), \quad \text{for } 0 < z < a.$$

II: *Crofton's Route* Let $F(a, v)$ be the distribution of V , and consider $F(a + h, v)$. By conditioning on the three events $\{\text{both } X \text{ and } Y \text{ lie in } (0, a)\}$, $\{\text{one of } X, Y \text{ lies in } (0, a)\}$, and $\{\text{neither of } X, Y \text{ lie in } (0, a)\}$, we find that

$$F(a + h, v) = F(a, v) \left(\frac{a}{a+h}\right)^2 + \left(\frac{v}{a} + o(h)\right) \frac{2ha}{(a+h)^2} + o(h).$$

Hence, rearranging and taking the limit as $h \rightarrow 0$, we have

$$(2) \quad \frac{\partial F}{\partial a}(a, v) = -\frac{2}{a}F(a, v) + \frac{2v}{a^2}.$$

Integrating (2) using the condition $F(a, a) = 1$ gives

$$(3) \quad F(a, v) = \frac{2av - v^2}{a^2}.$$

The densities of U and W may be found by the same method (exercise).

III: *Symmetry* Suppose we pick three points independently at random on the perimeter of a circle with perimeter of length a . Then choose any of the three as origin and “unwrap” the perimeter onto $(0, a)$. The other two points are distributed as X and Y . However, by the symmetry of the original problem the three lengths U, V , and W have the same density. By method I, part (i), it is $2(a - z)/a^2$.

(b) Let θ, ϕ, ψ be the angles subtended at the centre by the three sides of the triangle. The area of the triangle is $A = \frac{1}{2}r^2(\sin \theta + \sin \phi + \sin \psi)$; note that this expression is still valid when an angle is obtuse. However, by part (a), each of the arc lengths $r\theta, r\phi$, and $r\psi$ has the same density $2(2\pi r - z)/(2\pi r)^2$. Hence, θ has density $2(2\pi - \theta)/(2\pi)^2$, and

$$\mathbf{E}(A) = \frac{3}{2}r^2\mathbf{E}(\sin \theta) = \frac{3}{2}r^2 \int_0^{2\pi} \frac{2(2\pi - \theta) \sin \theta}{(2\pi)^2} d\theta = \frac{3r^2}{2}.$$

- (4) **Exercise** Show that the probability that U, V , and W can form a triangle is $\frac{1}{4}$.
- (5) **Exercise** Find the densities of U and W by method II.
- (6) **Exercise** Suppose that X_1, X_2, \dots, X_n are independently and uniformly distributed on $(0, 1)$ with order statistics $X_{(1)}, \dots, X_{(n)}$. What is the density of $X_{(k+1)} - X_{(k)}$; $1 \leq k \leq n - 1$?
- (7) **Exercise (6) Continued** What is the joint density of $X_{(k+1)} - X_{(k)}$ and $X_{(j+1)} - X_{(j)}$ for $j \neq k$?
- (8) **Exercise** Two points are picked at random on the perimeter (including its diameter) of a semicircle with radius 1. Show that the expected area of the resulting triangle they make with the midpoint of the diameter is $1/(2 + \pi)$.
- (9) **Exercise** Write down the joint density of U and W ; then integrate to derive (1) by a fourth method.

8.13 Example: Buffon’s Needle

An infinite horizontal table is marked with a rectangular grid comprising two families of distinct lines A and B . The lines of A are parallel, and the distance between neighbouring lines is $2a$. All the lines of B are perpendicular to every line of A and are distance $2b$ apart.

A thin symmetrical needle of length $2l$, where $l < \min\{a, b\}$, is thrown at random onto the table.

(a) Show that the probability that the needle intersects both an A -line and a B -line is

$$(1) \quad \mathbf{P}(A \cap B) = \frac{l^2}{\pi ab}.$$

(b) Show that the probability that the needle intersects an A -line and does not intersect a B -line is

$$(2) \quad \mathbf{P}(A \cap B^c) = \frac{2bl - l^2}{\pi ab}.$$

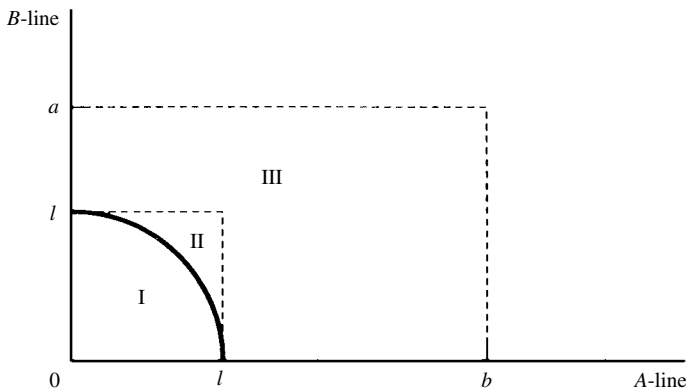


Figure 8.2 Buffon's needle.

Solution The centre C of the needle must fall in some $2a \times 2b$ rectangle R , whose sides are A -lines and B -lines. The words "at random" mean that the centre is uniformly distributed over R , and the angle Θ that the needle makes with any fixed line is uniformly distributed. By symmetry, we can suppose that C lies in one quarter of R , namely, the $a \times b$ rectangle Q , and also that $0 \leq \theta < \pi$. That is to say, we assume that $C = (X, Y)$ and Θ are jointly uniform on $\{0 \leq x < a\} \times \{0 \leq y < b\} \times \{0 \leq \theta < \pi\}$ with joint density $(\pi ab)^{-1}$.

(i) Now consider Figure 8.2. The needle can intersect both A and B only when $C = (x, y)$ lies in the positive quadrant of the circle, radius l , centred at the origin (region I).

If the angle it makes with OB lies between $\pm \sin^{-1}(x/l)$, then it cuts only OA . Likewise the needle cuts only OB if it lies within the angle $-\sin^{-1}(y/l) < \Theta < \sin^{-1}(y/l)$.

Therefore, when $X = x > 0$, $Y = y > 0$, and $x^2 + y^2 \leq l^2$, the probability of two intersections is $\pi - 2 \sin^{-1}(x/l) - 2 \sin^{-1}(y/l)$. Hence,

$$(3) \quad \mathbf{P}(A \cap B) = \frac{1}{\pi ab} \int \int \pi - 2 \sin^{-1} \left(\frac{x}{l} \right) - 2 \sin^{-1} \left(\frac{y}{l} \right) dx dy,$$

where the integral is over $x > 0$, $y > 0$, $x^2 + y^2 \leq l^2$. Now

$$\begin{aligned} \int_0^l \int_0^{\sqrt{l^2-x^2}} \sin^{-1} \left(\frac{x}{l} \right) dx dy &= \int_0^l (l^2 - x^2)^{\frac{1}{2}} \sin^{-1} \left(\frac{x}{l} \right) dx \\ &= \int_0^{\pi/2} l^2 \theta \cos^2 \theta d\theta, \quad \text{with the obvious substitution,} \\ &= \left(\frac{\pi^2}{16} - \frac{1}{4} \right) l^2. \end{aligned}$$

Hence, substituting into (3) gives

$$\mathbf{P}(A \cap B) = \frac{1}{\pi ab} \left(\frac{l^2 \pi^2}{4} - 4l^2 \left(\frac{\pi^2}{16} - \frac{1}{4} \right) \right) = \frac{l^2}{\pi ab}.$$

(ii) For $\mathbf{P}(A \cap B^c)$, we examine Figure 8.2 again. First, if C is in region I, then the needle cuts A and not B if Θ lies in an angle $2 \sin^{-1}(x/l)$, as we remarked above.

Second, if C lies in region II (that is, $0 \leq y < l$, but $x^2 + y^2 > l^2$), then the needle cuts A and not B if it lies in an angle of size $2 \cos^{-1}(y/l)$. Hence,

$$\begin{aligned} \pi ab \mathbf{P}(A \cap B^c) &= \int_{\text{I}} 2 \sin^{-1} \left(\frac{x}{l} \right) dx dy + \int_{\text{II}} 2 \cos^{-1} \left(\frac{y}{l} \right) dx dy \\ &= \int_0^l 2(l^2 - x^2)^{\frac{1}{2}} \sin^{-1} \left(\frac{x}{l} \right) dx \\ &\quad + \int_0^l 2(l - (l^2 - y^2)^{\frac{1}{2}}) \cos^{-1} \left(\frac{y}{l} \right) dy \\ &\quad + \int_0^l 2(a - l) \cos^{-1} \left(\frac{y}{l} \right) dy \\ &= l^2 \int_0^{\frac{\pi}{2}} 2\theta(\cos^2 \theta - \sin^2 \theta) d\theta + l \int_0^{\frac{\pi}{2}} (2l + 2(a - l))\theta \sin \theta d\theta \\ &= 2al - l^2, \quad \text{as required.} \end{aligned}$$

- (4) **Exercise** Show that the probability that the needle intersects no line of the grid is

$$1 - \frac{2l}{\pi b} - \frac{2l}{\pi a} + \frac{l^2}{\pi ab}.$$

(Do this in two ways, one of which is an integral.)

- (5) **Exercise** Suppose the table is marked with only one set of parallel lines, each distance $2a$ from its next neighbour. Show that the probability that a needle of length $2l < 2a$ intersects a line is $2l/\pi a$. (Do this two ways also.)
- (6) **Exercise** Consider the problem of Exercise 5 when $2l > 2a$. Show that the probability of an intersection is

$$\frac{2}{\pi} \cos^{-1} \left(\frac{a}{l} \right) + \frac{2l}{\pi a} \left(1 - \left(1 - \frac{a^2}{l^2} \right)^{\frac{1}{2}} \right).$$

- (7) **Exercise** Suppose (instead of a needle) you roll a penny of radius l on to the grid of A -lines and B -lines. What is the probability that when it topples over it intersects a line? When is this an even chance?

8.14 Example: Targets

- (a) Let $(X_i; 1 \leq i \leq 2n + 1)$ be independently and uniformly distributed over $(-1, 1)$, and let $Y_n = X_{(n+1)}$ so that Y_n is the sample median of the X_i . Find the density of Y_n , and hence evaluate the integral $\int_0^1 (1 - x^2)^n dx$.
- (b) Now n shots hit a circular target. The points of impact are independently and uniformly distributed over the circle. Let Z_n be the radius of the largest circle concentric with the target which includes no hit. Find $\mathbf{E}(Z_n)$.

Solution (a) First note that the uniform distribution on $(-1, 1)$ is $F(x) = \frac{1}{2}(1+x)$. Now let A_k be the event that $X_{(n+1)} = X_k$; this occurs of course if n of the X_i are greater than X_k , and the remaining n are less than X_k . Then

$$\begin{aligned} \mathbf{P}(Y_n \leq y) &= \sum_{k=1}^{2n+1} \mathbf{P}(Y_n \leq y \cap A_k) = (2n+1)\mathbf{P}(Y_n \leq y \cap A_1), \quad \text{by symmetry,} \\ &= (2n+1) \int_{-1}^y f_{X_1}(y) \binom{2n}{n} \left(\frac{1+y}{2}\right)^n \left(1 - \frac{1+y}{2}\right)^n dy, \end{aligned}$$

by conditional probability.

Hence, Y_n has density $f_Y(y) = ((2n+1)!/(n!)^2)((1-y^2)^n/2^{2n+1})$. Because this is a density, its integral over $(-1, 1)$ is unity, so

$$\int_0^1 (1-y^2)^n dy = \frac{2^{2n}(n!)^2}{(2n+1)!}.$$

[Alternatively, you could write down the density for $f_Y(y)$ using the known density of order statistics.]

(b) Let R_i be the distance of the i th hit from the centre of the target. Because hits are uniform, $\mathbf{P}(R_i \leq x) = x^2$ for $0 \leq x \leq 1$. Obviously,

$$\begin{aligned} \mathbf{P}(Z_n > x) &= \mathbf{P}(R_i > x \text{ for all } i) = \prod_{i=1}^n \mathbf{P}(R_i > x), \quad \text{by independence,} \\ &= (1-x^2)^n. \end{aligned}$$

Hence,

$$\mathbf{E}(Z_n) = \int_0^1 \mathbf{P}(Z_n > x) dx = \int_0^1 (1-x^2)^n dx = \frac{2^{2n}(n!)^2}{(2n+1)!}.$$

- (1) **Exercise** Find $\text{var}(Y_n)$.
- (2) **Exercise** Let A_n be the area of the smallest circle concentric with the target that includes all the hits. Find $\mathbf{E}(A_n)$.
- (3) **Exercise** The hit furthest from the centre of the target is deleted. What now is the expected area of the smallest circle concentric with the target that includes all the remaining hits?
- (4) **Exercise** Let R_n be the distance of the furthest hit from the centre of the target. Show that as $n \rightarrow \infty$, $\mathbf{P}(n(1-R_n) \leq x) \rightarrow 1 - e^{-2x}$.

8.15 Example: Gamma Densities

Let X and Y be independent, having gamma distributions with parameters $\{\alpha, \lambda\}$ and $\{\beta, \lambda\}$, respectively.

- (a) Find the joint density of $U = X + Y$ and $V = \frac{X}{X+Y}$.
- (b) Deduce that $\mathbf{E}\left(\frac{X}{X+Y}\right) = \frac{\mathbf{E}(X)}{\mathbf{E}(X) + \mathbf{E}(Y)}$.
- (c) What is the density of V ?

Solution (a) We use the change of variables technique. The transformation $u = x + y$, $v = x/(x + y)$ for $x, y > 0$, is a one-one map of the positive quadrant onto the strip $0 < v < 1, u > 0$, with inverse $x = uv$ and $y = u(1 - v)$. Hence, $J = u$, and by Theorem 8.2.1, U and V have joint density

$$(1) \quad f(u, v) = \frac{\lambda^\alpha \lambda^\beta}{\Gamma(\alpha)\Gamma(\beta)} (uv)^{\alpha-1} (u(1-v))^{\beta-1} e^{-\lambda uv} e^{-\lambda u(1-v)} u$$

$$(2) \quad = c_1 u^{\alpha+\beta-1} e^{-\lambda u} c_2 v^{\alpha-1} (1-v)^{\beta-1},$$

where c_1 and c_2 are constants.

Hence, U and V are independent, as $f(u, v)$ has factorized.

(b) Using the independence of U and V gives

$$(3) \quad \mathbf{E}(X) = \mathbf{E}(UV) = \mathbf{E}(U)\mathbf{E}(V) = (\mathbf{E}(X) + \mathbf{E}(Y))\mathbf{E}\left(\frac{X}{X+Y}\right),$$

as required.

(c) A glance at (2) shows that V has the beta density with parameters α and β .

$$(4) \quad \textbf{Exercise} \quad \text{Show that } \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

$$(5) \quad \textbf{Exercise} \quad \text{Show that } \Gamma\left(\frac{1}{2}\right) = \pi^{\frac{1}{2}}.$$

$$(6) \quad \textbf{Exercise} \quad \text{Let the random variable } Z \text{ have density } c(1+x^2)^{-m}, m > \frac{1}{2}, -\infty < x < \infty. \text{ Show that } c^{-1} = \pi^{\frac{1}{2}}\Gamma(m - \frac{1}{2})/\Gamma(m).$$

8.16 Example: Simulation – The Rejection Method

(a) Let U and X be independent random variables such that U is uniform on $(0, 1)$ and X has density $f_X(x)$; suppose that there exists a constant a that for all x the function $f_S(x)$ satisfies

$$(1) \quad 0 \leq f_S(x) \leq a f_X(x)$$

and $\int_{-\infty}^{\infty} f_S(x) dx = 1$. Show that

$$(2) \quad \mathbf{P}(X \leq x | aU f_X(X) \leq f_S(X)) = \int_{-\infty}^x f_S(y) dy.$$

(b) Explain how this result may be used to produce realizations of a random variable Z with density $f_S(z)$.

Solution (a) By conditional probability,

$$(3) \quad \mathbf{P}(X \leq x | aU f_X(X) \leq f_S(X)) = \frac{\mathbf{P}(X \leq x, aU f_X(X) \leq f_S(X))}{\mathbf{P}(aU f_X(X) \leq f_S(X))}$$

$$= \frac{\int_{-\infty}^x \mathbf{P}(aU f_X(x) \leq f_S(x)) f_X(x) dx}{\int_{-\infty}^{\infty} \mathbf{P}(aU f_X(x) \leq f_S(x)) f_X(x) dx}$$

$$\begin{aligned}
 &= \int_{-\infty}^x \frac{f_S(x)}{af_X(x)} f_X(x) dx \bigg/ \int_{-\infty}^{\infty} \frac{f_S(x)}{af_X(x)} f_X(x) dx, \\
 & \hspace{20em} \text{by (1),} \\
 &= \int_{-\infty}^x f_S(x) dx.
 \end{aligned}$$

(b) Suppose we have independent realizations of U and X . Then the above equation says that conditional on the event $A = \{aUf_X(X) \leq f_S(X)\}$, X has density $f_S(x)$. In familiar notation, we have

$$f_{X|A}(x) = f_S(x).$$

Now suppose we have a sequence $(U_k, X_k; k \geq 1)$ of random variables that have the same distributions as (U, X) . For every pair (U_k, X_k) for which A occurs, the random variable X_k has density $f_S(x)$, and we can write $Z = X_k$. Then Z has density $f_S(z)$.

Remark It is implicit in the question that we want a random variable with density $f_S(x)$, and so any pair (U_k, X_k) for which A^c occurs is rejected. This explains the title of the example (although in the circumstances you might think a better title would be the *conditional* method). Obviously, this offers a method for simulating random variables with an arbitrary density $f_S(x)$, subject only to the constraint that we have to be able to simulate X with density $f_X(x)$ that satisfies (1).

- (4) **Exercise** Find the mass function and mean of the number N of pairs (U_k, X_k) that are rejected before the first occasion on which A occurs. What does this imply about a ?
- (5) **Exercise** If X is exponential with parameter 1, show that (2) takes the form $\mathbf{P}(X \leq x | aU_1U_2 \leq f_S(X)) = F_S(x)$, where U_1 and U_2 are independent and uniform on $(0, 1)$. Hence, describe how you would simulate a random variable with density $f_S(x) = (2/\pi)^{1/2} e^{-x^2/2}$, $x > 0$.
- (6) **Exercise** Let U_1 and U_2 be independent and uniform on $(0, 1)$. Let $X = -\log U_1$ and $Y = -\log U_2$. What is the density of X conditional on $Y > \frac{1}{2}(X - 1)^2$?

8.17 Example: The Inspection Paradox

Let $N(t)$ be a Poisson process, and at each time $t > 0$, define $C(t)$ to be the time since the most recent event. (This is called the *current life* or *age*.) Further, define $B(t)$ to be the time until the next event (this is called the *balance of life* or *excess life*). Show that $B(t)$ and $C(t)$ are independent, and find the distribution of $C(t)$. What is $\mathbf{E}(B + C)$?

[Note: By convention, if $N(t) = 0$, we set $C(t) = t$.]

Solution Recall that we used the conditional property of the Poisson process to show that $N(t)$ has independent increments. Now

$$\begin{aligned}
 \mathbf{P}(B(t) > y, C(t) > z) &= \mathbf{P}(N(t+y) - N(t) = 0, N(t) - N(t-z) = 0) \\
 &= \mathbf{P}(N(t+y) - N(t) = 0) \mathbf{P}(N(t) - N(t-z) = 0), \\
 & \hspace{10em} \text{by the independence of increments,} \\
 &= \mathbf{P}(B(t) > y) \mathbf{P}(C(t) > z).
 \end{aligned}$$

Furthermore, we showed that $N(t) - N(t - z)$ has the same distribution as $N(z)$, for t and $t - z$ both nonnegative. Hence,

$$(1) \quad \mathbf{P}(C(t) > z) = \begin{cases} 1 & z < 0 \\ e^{-\lambda z} & 0 \leq z \leq t \\ 0 & z > t. \end{cases}$$

Likewise,

$$\mathbf{P}(B(t) > y) = \begin{cases} 1 & y < 0 \\ e^{-\lambda y} & y > 0. \end{cases}$$

Hence,

$$(2) \quad \mathbf{E}(B + C) = \frac{1}{\lambda} + \int_0^t \lambda t e^{-\lambda t} dt + t e^{-\lambda t} = \frac{2}{\lambda} - \frac{1}{\lambda} e^{-\lambda t}.$$

Remark If we suppose $N(t)$ is the number of renewals of (say) light bulbs, then (2) says that the expected life of the light bulb inspected at time t is $2/\lambda - 1/\lambda e^{-\lambda t}$, which is greater than the expected life of a randomly selected light bulb, which is $1/\lambda$.

It may seem as though we make light bulbs last longer by inspecting them, this is the “paradox.” Of course, this is not so, it is just that if you only look once, you are more likely to see a longer-lived light bulb. This is related to other sampling paradoxes mentioned previously, see for example, “congregations.”

- (3) **Exercise: The Markov Property** Show that for any $t_1 < t_2 < \dots < t_n$, the process $N(t)$ has the so-called Markov property:

$$\mathbf{P}(N(t_n) = j_n | N(t_{n-1}) = j_{n-1}, \dots, N(t_1) = j_1) = \mathbf{P}(N(t_n) = j_n | N(t_{n-1}) = j_{n-1}).$$

- (4) **Exercise: The Shower Problem** Your telephone is called at the instants of a Poisson process with parameter λ . Each day you take a shower of duration Y starting at time X , where X and Y are jointly distributed in hours (and not independent). Show that the number of times that the telephone is called while you are in the shower has a Poisson distribution with parameter $\lambda \mathbf{E}(Y)$. (Assume $0 \leq X \leq X + Y \leq 24$.)
- (5) **Exercise** Aesthetes arrive at a small art gallery at the instants of a Poisson process of parameter λ . The k th arrival spends a time X_k in the first room and Y_k in the second room, and then leaves. The random variables X_k and Y_k are not independent, but (X_k, Y_k) is independent of (X_j, Y_j) for $j \neq k$. At time t , let R_1 and R_2 be the number of aesthetes in the respective rooms. Show that R_1 and R_2 are independent Poisson random variables.
- (6) **Exercise** Find $\text{cov}(N(s), N(t))$, and the correlation $\rho(N(s), N(t))$

8.18 Example: von Neumann’s Exponential Variable

Let the sequence of random variables X_1, X_2, X_3, \dots be independent and identically distributed with density f and distribution F . Define the random variable R by

$$(1) \quad R = \min\{n : X_1 \geq X_2 \geq \dots \geq X_{n-1} < X_n\}.$$

(a) Show that $\mathbf{P}(R = r) = (r - 1)/r!$, and that

$$\mathbf{P}(X_R \leq x) = \exp(1 - F(x)) - e(1 - F(x)).$$

(b) Now let X_n be uniformly distributed on $(0, 1)$ for all n . Show that

$$(2) \quad \mathbf{P}(X_1 \leq x; R = r) = \frac{x^{r-1}}{(r-1)!} - \frac{x^r}{r!}.$$

Deduce that

$$(3) \quad \mathbf{P}(X_1 \leq x | R \text{ is even}) = \frac{1 - e^{-x}}{1 - e^{-1}}.$$

Finally, define a random variable V as follows. A sequence X_1, X_2, \dots, X_R is a “run”; it is “odd” if R is odd, otherwise it is “even.” Generate runs until the first even run, and then let V equal the number N of odd runs plus X_1 in the even run. Show that V has density e^{-v} for $v > 0$.

Solution (a) Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(r)}$ be the order statistics of (X_1, \dots, X_r) . By symmetry (X_1, \dots, X_r) is equally likely to be any one of the $r!$ permutations of $(X_{(1)}, \dots, X_{(r)})$. For the $r - 1$ permutations of the form,

$$(X_{(r)}, \dots, X_{(k+1)}, X_{(k-1)}, \dots, X_{(1)}, X_{(k)}), \quad 2 \leq k \leq r,$$

the event $R = r$ occurs, and for no others. Hence,

$$\mathbf{P}(R = r) = \frac{r - 1}{r!}.$$

The above remarks also show that

$$\begin{aligned} \mathbf{P}(X_R \leq x) &= \sum_{r=2}^{\infty} \frac{1}{r!} \sum_{k=2}^r \mathbf{P}(X_{(k)} \leq x) \\ &= \sum_{r=2}^{\infty} \frac{1}{r!} \sum_{k=2}^r \sum_{j=k}^r \binom{r}{j} (F(x))^j (1 - F(x))^{r-j} \quad \text{by (8.7.12)} \\ &= \sum_{r=2}^{\infty} \frac{1}{r!} (rF(x) - 1 + (1 - F(x))^r) = e(F(x) - 1) + \exp(1 - F(x)), \end{aligned}$$

on summing the series. It is easy to check that this is continuous and nondecreasing as x increases, and differentiation gives the density of X_R :

$$f_{X_R}(x) = e f(x)(1 - \exp(1 - F(x))).$$

(b) Now observe that the event $\{R > r\} \cap \{X_1 \leq x\}$ occurs if and only if $X_k \leq x$ for $1 \leq k \leq r$ [which has probability $(F(x))^r$, and $X_1 \geq X_2 \geq X_3 \geq \dots \geq X_r$ (which has probability $1/r!$)]. Hence, when X_k is uniform on $(0, 1)$,

$$\mathbf{P}(R > r; X_1 \leq x) = \frac{x^r}{r!}$$

and so

$$\mathbf{P}(R = r; X_1 \leq x) = \frac{x^{r-1}}{(r-1)!} - \frac{x^r}{r!}.$$

Now summing over even values of R , we have

$$\mathbf{P}(X_1 \leq x, R \text{ is even}) = 1 - e^{-x},$$

and hence,

$$\mathbf{P}(X_1 \leq x | R \text{ is even}) = (1 - e^{-x}) / (1 - e^{-1}).$$

This shows that $\mathbf{P}(R \text{ is even}) = 1 - e^{-1}$, and so by independence of runs, N is a geometric random variable, with mass function

$$\mathbf{P}(N = n) = e^{-n}(1 - e^{-1}), \quad \text{for } n > 0.$$

Finally, let us denote X_1 in the even run by X_0 . Then $\mathbf{P}(X_0 > x) = 1 - (1 - e^{-x}) / (1 - e^{-1})$, from (3). Hence,

$$\begin{aligned} (4) \quad \mathbf{P}(V > v) &= \mathbf{P}(N \geq [v] + 1) + \mathbf{P}(N = [v]; X_0 > v - [v]) \\ &= e^{-[v]-1} + ((1 - e^{-1})e^{-[v]}) \left(1 - \frac{1 - e^{-v+[v]}}{1 - e^{-1}}\right) \\ &= e^{-v} \quad \text{for } 0 < v < \infty. \end{aligned}$$

Thus, V is exponentially distributed.

Remark This method of generating exponential random variables from uniform ones was devised by von Neumann in 1951. Notice that it is computationally economical, in that it is necessary to store only the number of odd runs to date, and the first X_1 in the run in progress. Also, the expected number of uniform random variables used for each exponential random variable is small. Since the original result, the method has been extended to generate other continuous random variables from uniform r.v.s.

- (5) **Exercise** What is the density of X_{R-1} ?
- (6) **Exercise: Bad luck** As above, X_1, X_2, \dots are independent and identically distributed with density f and distribution F . Define $T = \min\{n: X_n > X_1\}$. Find $\mathbf{P}(T = n)$ and show that T has infinite expectation. Show that X_T has distribution $F_{X_T}(x) = F(x) + (1 - F(x)) \log(1 - F(x))$.
- (7) **Exercise** Explain why the above exercise is entitled “Bad luck.”
- (8) **Exercise** Use the result of Exercise 6 to show that when X_1 has an exponential distribution, X_T has a gamma distribution. Why is this obvious without going through the analysis of Exercise 6?

8.19 Example: Maximum from Minima

Let X_1, \dots, X_n be a collection of nonnegative random variables with finite expected values. Show that

$$\begin{aligned} \mathbf{E} \max_j X_j &= \sum_j \mathbf{E} X_j - \sum_{j < k} \mathbf{E} \min(X_j, X_k) \\ &\quad + \sum_{i < j < k} \mathbf{E} \min(X_i, X_j, X_k) - \dots + (-1)^{n+1} \mathbf{E} \min_j X_j. \end{aligned}$$

Solution Let $I_j = I(X_j > x)$ be the indicator of the event that $X_j > x$. Then for all x we have $I_j I_k = I(\min(X_j, X_k) > x)$, and so on for any product. Furthermore, for all x , we see by inspection that

$$\begin{aligned} I(\max I_j > x) &= 1 - \prod_{j=1}^n (1 - I_j) \\ &= \sum_j I(X_j > x) - \sum_{j < k} I(X_j \wedge X_k > x) + \cdots \\ &\quad + (-1)^{n+1} I(\min X_j > x). \end{aligned}$$

Now taking expectations gives

$$\mathbf{P}(\max X_j > x) = \sum_j \mathbf{P}(X_j > x) - \sum_{j < k} \mathbf{P}(X_j \wedge X_k > x) + \cdots$$

Finally, integrating over $[0, \infty)$ and recalling the tail integral for expectation, Theorem 7.4.11, gives the result.

- (1) **Exercise** Deduce that for any collection of real numbers (x_1, \dots, x_n) we have

$$\max_j x_j = \sum_j x_j - \sum_{j < k} x_j \wedge x_k + \cdots + (-1)^{n+1} \min_j x_j.$$

- (2) **Exercise: Coupon Collecting (Example 5.3.3) Revisited** Suppose that at each transaction you acquire a coupon of the i th type with probability c_i , $1 \leq i \leq n$, where $\sum_i c_i = 1$ and $c_i \neq 0$. Show that the expected number of coupons collected until you first have at least one of every type is

$$\sum_j \frac{1}{c_j} - \sum_{j < k} \frac{1}{c_j + c_k} + \sum_{i < j < k} \frac{1}{c_i + c_j + c_k} - \cdots + (-1)^n (-1)^{n+1} \frac{1}{\sum_i c_i}.$$

- (3) **Exercise** A wood contains n birds of r species, with b_i of the i th species, $b_i \neq 0$, so that $\sum_{i=1}^r b_i = n$. You catch and cage birds at random one by one. Let B be the expected number of captures until you have at least one of every species, and C the expected number of captures until you have caged all the representatives of any one of the species. Show that

$$\frac{B}{n+1} = \sum_j \frac{1}{b_j + 1} - \sum_{j < k} \frac{1}{b_j + b_k + 1} + \cdots + \frac{(-)^{r+1}}{b_1 + \cdots + b_r + 1},$$

and $C = n - B + 1$.

- (4) **Exercise** A complex machine contains r distinct components, and the lifetime to failure of the i th component is an exponential random variable with parameter a_i , independent of all the rest. On failing, it is replaced by one with identical and independent properties. Find an expression for the expected time until every distinct component has failed at least once.

8.20 Example: Binormal and Trinormal

Let U and V be independent random variables having the standard normal $N(0, 1)$ density. Set

$$(1) \quad X = U, \quad Y = \rho U + \sqrt{1 - \rho^2}V; \quad |\rho| < 1.$$

Show that the pair X, Y has the standard bivariate normal density. Deduce that

$$\mathbf{P}(X > 0, Y > 0) = \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \rho.$$

Solution The joint density of U and V is

$$f(u, v) = (2\pi)^{-1} \exp \left\{ -\frac{1}{2}(u^2 + v^2) \right\},$$

and the inverse transformation is

$$U = X, \quad V = (Y - \rho X)/(1 - \rho^2)^{1/2}.$$

Hence, by Theorem 8.2.1, X and Y have joint density

$$f(u(x, y), v(x, y))|J| = \frac{1}{2\pi(1 - \rho^2)^{1/2}} \exp \left\{ -\frac{1}{2(1 - \rho^2)}(x^2 - 2\rho xy + y^2) \right\},$$

which is indeed the standard bivariate normal density. Using this, and recalling Example 8.3.8, we find

$$\begin{aligned} \mathbf{P}(X > 0, Y > 0) &= \mathbf{P}(U > 0, \rho U + \sqrt{1 - \rho^2}V > 0) \\ &= \mathbf{P}\left(\frac{\rho}{(1 - \rho^2)^{1/2}} < \tan^{-1} \frac{V}{U} < \infty\right) \\ &= \left\{ \frac{\pi}{2} + \tan^{-1} \frac{\rho}{(1 - \rho^2)^{1/2}} \right\} \cdot \frac{1}{2\pi} \\ &= \frac{1}{4} + \frac{1}{2\pi} \sin^{-1} \rho, \end{aligned}$$

where we use the well-known fact that

$$\sin \theta = \rho \Leftrightarrow \cos \theta = \sqrt{1 - \rho^2} \Leftrightarrow \tan \theta = \frac{\rho}{\sqrt{1 - \rho^2}}.$$

- (2) **Exercise** Show that $aX + bY$ has the normal density $N(0, a^2 + 2\rho ab + b^2)$.
- (3) **Exercise** Show that X and Y are independent if and only if $\text{cov}(X, Y) = 0$.
- (4) **Exercise** Show that

$$\mathbf{E}(Y|aX + bY = z) = \frac{b^2 + \rho ab}{a^2 + 2\rho ab + b^2} \frac{z}{b}$$

and

$$\text{var}(Y|aX + bY = z) = \frac{a^2(1 - \rho^2)}{a^2 + 2ab + b^2}.$$

- (5) **Exercise** Let $X, Y,$ and Z have the standard trivariate normal density, defined in (8.10) Show that

$$\mathbf{P}(X > 0, Y > 0, Z > 0) = \frac{1}{8} + \frac{1}{4\pi} \left\{ \sin^{-1} \rho_{12} + \sin^{-1} \rho_{23} + \sin^{-1} \rho_{31} \right\}.$$

Express $X, Y,$ and Z in terms of three independent $N(0, 1)$ random variables $U, V,$ and $W.$ Hence, deduce that

$$\mathbf{E}(Z|X, Y) = \{(\rho_{31} - \rho_{12}\rho_{23})X + (\rho_{23} - \rho_{12}\rho_{31})Y\} / (1 - \rho_{12}^2)$$

and

$$\text{var}(Z|X, Y) = \{1 - \rho_{12}^2 - \rho_{23}^2 - \rho_{31}^2 + 2\rho_{12}\rho_{23}\rho_{31}\} / (1 - \rho_{12}^2).$$

- (6) **Exercise** Do Example 8.11 again, this time using the representation in (1) for X and $Y.$

8.21 Example: Central Limit Theorem

Let $(X_n; n \geq 1)$ be a collection of independent Poisson random variables with parameter 1. By applying the central limit theorem to the $X_n,$ prove Stirling's formula:

$$\lim_{n \rightarrow \infty} \sqrt{n} e^{-n} n^n / n! = (2\pi)^{-\frac{1}{2}}.$$

[You may assume without proof that the convergence of the sequence of distributions to $\Phi(x)$ is uniform in x on finite intervals including 0.]

Solution Let $S_n = \sum_{r=1}^n X_r.$ Then S_n is Poisson with parameter $n,$ mean $n,$ and variance $n.$ Thus, $\mathbf{P}(S_n = n) = e^{-n} n^n / n!,$ and we may write

$$\begin{aligned} (1) \quad \sqrt{n} e^{-n} n^n / n! &= \sqrt{n} \mathbf{P}(S_n = n) = \sqrt{n} \mathbf{P}(n-1 < S_n \leq n) \\ &= \sqrt{n} \mathbf{P}\left(\frac{n-1-n}{\sqrt{n}} < \frac{S_n-n}{\sqrt{n}} \leq 0\right) \\ &= \sqrt{n} \left[F_n(0) - F_n\left(-\frac{1}{\sqrt{n}}\right) \right], \end{aligned}$$

where F_n is the distribution function of $(S_n - n)/\sqrt{n},$ and by the Central Limit Theorem, we have

$$(2) \quad F_n(x) \rightarrow \Phi(x) \quad \text{as } n \rightarrow \infty.$$

Furthermore, because $\Phi'(x) = \phi(x),$ we have, as $n \rightarrow \infty,$

$$(3) \quad \sqrt{n} \left\{ \Phi(0) - \Phi\left(-\frac{1}{\sqrt{n}}\right) \right\} \rightarrow \phi(0) = \frac{1}{\sqrt{2\pi}}.$$

Because the convergence in (2) is uniform on finite intervals, we may let $n \rightarrow \infty$ in (1) and use (3) to yield the result.

- (4) **Exercise** Apply the central limit theorem to the same family $(X_n; n \geq 1)$ of Poisson random variables to show that

$$\lim_{n \rightarrow \infty} e^{-n} \left(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!} \right) = \frac{1}{2}.$$

- (5) **Exercise** It is said that D. Hagelbarger built a machine to predict whether a human coin-flipper would call heads or tails. In 9795 flips, the machine was correct on 5218 occasions. What is the probability of doing at least this well by chance? (The flipped coin was known to be fair by all involved.)
- (6) **Exercise** An aeroplane has 120 seats and is full. There are 120 inflight meals, of which 60 are fish and 60 are pasta. Any passenger, independently of the rest, prefers pasta with probability 0.55, or prefers fish with probability 0.45. Show that the probability that 10 or more passengers will not get their first choice is approximately 0.234. [You are given that $\Phi(0.734) \simeq 0.7676$ and $\Phi(2.94) \simeq 0.9984$.]

8.22 Example: Poisson Martingales

Suppose that $(N(t); t \geq 0)$ is a collection of nonnegative integer-valued random variables such that $N(s) \leq N(t)$ for all $0 \leq s \leq t < \infty$, and $W(t) = \exp\{-\theta N(t) + \lambda t(1 - e^{-\theta})\}$ is a martingale. Show that $N(t)$ is a Poisson process.

Solution For $s < t$,

$$\begin{aligned} & \mathbf{E}(\exp(-\theta(N(t) - N(s)) | N(u); 0 \leq u \leq s) \\ &= \mathbf{E}\left(\frac{W(t)}{W(s)} \exp[(-\lambda(t - s))(1 - e^{-\theta})] | N(u); 0 \leq u \leq s\right) \\ &= \exp[-\lambda(t - s)(1 - e^{-\theta})] \end{aligned}$$

because $W(t)$ is a martingale. As this does not depend on $N(u)$, $0 \leq u \leq s$, it follows that $N(t)$ has independent increments. Furthermore, we recognise the final expression as the moment generating function of a Poisson random variable with parameter $\lambda(t - s)$. Hence, $N(t)$ is a Poisson process.

- (1) **Exercise** Let $N(t)$ be a Poisson process with parameter λ . Show that $W(t)$ defined above is a martingale.
- (2) **Exercise** Let $N(t)$ be a Poisson process with parameter λ , and $N(0) = 0$. Let $T = \min\{t: N(t) = a\}$, where a is a positive integer. Use the optional stopping theorem to show that
- (a) $\mathbf{E}T = a/\lambda$.
- (b) $\text{var}T = a/\lambda^2$.
- (c) $\mathbf{E}e^{-\theta T} = \left(\frac{\lambda}{\lambda + \theta}\right)^a$.
- (Hint. Recall the martingales of Example 8.8.17.)
- (3) **Exercise: Integrated Poisson process.** Let $N(t)$ be a Poisson process with parameter λ . Show that

$$\int_0^t N(u) du - \frac{1}{2\lambda} N(t)^2 + \frac{1}{2\lambda} N(t)$$

is a martingale. If $T = \min\{t: N(t) = a\}$ where a is a positive integer, deduce that

$$\mathbf{E} \int_0^T N(u) du = \frac{1}{2\lambda} a(a - 1).$$

8.23 Example: Uniform on the Unit Cube

Let X , Y , and Z be independent, identically distributed, and uniform on $[0, 1]$. Show that $W = (XY)^Z$ is also uniform on $(0, 1)$.

Solution First, we recall from Example 7.2.2 that the random variable U has a uniform density on $(0, 1)$ if and only if $-\log U$ has an exponential density on $(0, \infty)$. Hence, taking logarithms and using Example 8.5.15 and Theorem 8.6.11, we have

$$\begin{aligned} \mathbf{E} \exp[-t \log W] &= \mathbf{E} \exp\{tZ\{-\log X - \log Y\}\} \\ &= \mathbf{E}\{\mathbf{E}\{\exp[tZ(-\log X - \log Y)]|Z\}\} \\ &= \mathbf{E} \frac{1}{(1-tZ)^2}, \quad \text{since } -\log X \text{ and } -\log Y \text{ are exponential,} \\ &= \int_0^1 \frac{1}{(1-tz)^2} dz = \frac{1}{1-t}. \end{aligned}$$

Hence, by Example 7.5.4, $-\log W$ has an exponential density, so that W has a uniform density on $(0, 1)$, by the remark above.

Remark In the following exercises, X , Y , and Z are independent and uniform on $(0, 1)$, with order statistics $X_{(1)}, X_{(2)}, X_{(3)}$.

- (1) **Exercise** Let $S = X + Y + Z$. Show that S has density

$$f_S(s) = \begin{cases} s^2/2, & 0 \leq s < 1 \\ -s^2 + 3s - 3/2, & 1 \leq s < 2 \\ s^2/2 - 3s + 9/2, & 2 \leq s \leq 3. \end{cases}$$

- (2) **Exercise** Regarding X , Y , and Z as three points on the unit interval, show that the probability that the distance between any two is at least d is $(1 - 2d)^3$, when $d < \frac{1}{2}$.
- (3) **Exercise** Show that the probability that $Xt^2 + Yt + Z = 0$ has two real roots is $\frac{5}{36} + \frac{1}{6} \log 2$.
- (4) **Exercise** Show that the probability that $Xt^3 - 3Yt + Z = 0$ has three real roots is $1 - \frac{5}{4}r^{-1} + \frac{2}{5}\sqrt{r}$, where $r = 4^{1/3}$.
- (5) **Exercise** Show that the joint density of $V = \frac{X_{(1)}}{X_{(2)}}$ and $W = \frac{X_{(2)}}{X_{(3)}}$ is $f(v, w) = 2w$ for $0 \leq v, w \leq 1$. Deduce that V and W are independent.

8.24 Example: Characteristic Functions

For any random variable X , define the function $\phi(t) = \mathbf{E}e^{itX}$, where $i = \sqrt{-1}$. Show that for all real t

$$|\phi(t)| \leq 1,$$

and find $\phi(t)$ when X is exponential with parameter λ .

Solution By definition of e^{itX} ,

$$\mathbf{E}e^{itX} = \mathbf{E}(\cos tX + i \sin tX),$$

but

$$|\cos tX + i \sin tX| = \cos^2 tX + \sin^2 tX = 1,$$

and so $|\phi(t)| \leq 1$. If X is exponential,

$$\mathbf{E}e^{itX} = \int_0^\infty \lambda e^{-\lambda x} \cos tx \, dx + i \int_0^\infty \lambda e^{-\lambda x} \sin tx \, dx.$$

Integrating by parts gives

$$\int_0^\infty \lambda e^{-\lambda x} \cos tx \, dx = 1 - \int_0^\infty t e^{-\lambda x} \sin tx \, dx = 1 - \int_0^\infty \frac{t^2}{\lambda^2} \lambda e^{-\lambda x} \cos tx \, dx.$$

Hence,

$$\int_0^\infty \lambda e^{-\lambda x} \cos tx \, dx \left(1 - \frac{t^2}{\lambda^2}\right) = 1$$

Likewise,

$$\int_0^\infty \lambda e^{-\lambda x} \sin tx \, dx \left(1 + \frac{t^2}{\lambda^2}\right) = \frac{t}{\lambda}.$$

Hence,

$$\mathbf{E}e^{itX} = \frac{1 + it/\lambda}{1 + t^2/\lambda^2} = \frac{\lambda}{\lambda - it}.$$

Remark $\phi(t)$ is called the *characteristic function* of X . We have occasionally been frustrated by the nonexistence of a moment generating function. This example shows that we can use the characteristic function in such cases because it always exists. For example, it can be shown that the Cauchy density $f(x) = \{\pi(1+x^2)\}^{-1}$, which has no moments at all, has characteristic function $\phi(t) = e^{-|t|}$.

Furthermore, for any random variable X that does have a moment generating function $M(\theta) = \mathbf{E}e^{\theta X}$, for $|\theta| < \varepsilon > 0$, the characteristic function of X is given by $\phi(t) = M(it)$.

- (1) **Exercise** Show that the random variable uniform on $[-1, 1]$ has characteristic function $\phi(t) = \frac{\sin t}{t}$.
- (2) **Exercise** If X_1, \dots, X_n are independent Cauchy random variables, show that $\bar{X} = \frac{1}{n} \sum_1^n X_i$ has the same Cauchy density as the X_i .
- (3) **Exercise** Find the characteristic function of the random variable X with density

$$f(x) = \frac{1}{2}e^{-|x|}, \quad x \in \mathbb{R}.$$

PROBLEMS

- 1 When is $f(x, y) = xy + ax + by + 1$ a joint density function on $0 \leq x, y \leq 1$? Can it be the joint density of random variables X and Y that are independent?
- 2 Find $\text{cov}(X, Y)$ for the joint density of Problem 1.

- 3 Let X and Y have joint density $f(x, y) = c \exp(-x - y)$ for $x > 0, y > 0$. Find (a) c , (b) $\mathbf{P}(X + Y > 1)$, and (c) $\mathbf{P}(X < Y)$.
- 4 Let X and Y have joint density $f(x, y) = g(x + y)$ for $x \geq 0, y \geq 0$. Find the density of $Z = X + Y$.
- 5 Let X and Y have joint density $f = c(1 + x^2 + y^2)^{-\frac{3}{2}}$ for all x and y . (a) What is c ? (b) Find the marginal density of X .
- 6 Let X and Y have the joint density of Problem 5, and define $W = X^2 + Y^2, Z = Y/X$. Show that W and Z are independent.
- 7 Let U and V be independently and uniformly distributed on $(0, 1)$. Find the joint density of $X = U^{\frac{1}{2}}/(U^{\frac{1}{2}} + V^{\frac{1}{2}})$ and $Y = U^{\frac{1}{2}} + V^{\frac{1}{2}}$. By considering $\mathbf{P}(X \leq x|Y \leq 1)$, devise a rejection sampling procedure for simulating a random variable with density $6x(1 - x); 0 < x < 1$.
- 8 Let U_1, U_2 , and U_3 be independently and uniformly distributed on $(0, 1)$, with order statistics $U_{(1)} < U_{(2)} < U_{(3)}$. Show that the density of $U_{(2)}$ is $6x(1 - x); 0 < x < 1$.
- 9 Let U_1, U_2, U_3 , and U_4 be independently and uniformly distributed on $(0, 1)$. What is the density of $X = \log(U_1 U_2)/\log(U_1 U_2 U_3 U_4)$? (*Hint*: One way uses Example 8.15).
- 10 Let X and Y be independent normal random variables, and set $U = X + Y, V = X - Y$. Show that U and V are independent if and only if $\text{var}(X) = \text{var}(Y)$.
- 11 **Simulation Using Bivariate Rejection** Let U and V be independent and uniform on $(0, 1)$. Define the random variables

$$\begin{aligned} Z &= (2U - 1)^2 + (2V - 1)^2, \\ X &= (2U - 1)(2Z^{-1} \log Z^{-1})^{\frac{1}{2}} \\ Y &= (2V - 1)(2Z^{-1} \log Z^{-1})^{\frac{1}{2}}. \end{aligned}$$

Show that the conditional joint density of X and Y given $Z < 1$, is $(2\pi)^{-1} \exp(-\frac{1}{2}(x^2 + y^2))$. Explain how this provides a method for simulating normal random variables.

- 12 Let X and Y be independent exponential random variables with respective parameters λ and μ . Find $\mathbf{P}(\max\{X, Y\} \leq aX)$ for $a > 0$.
- 13 Let X and Y have joint density $f = cye^{-y(x+1)}, 0 \leq x < y < \infty$ for some constant c . What is the conditional density of X given $Y = y$?
- 14 A spherical melon has radius 1. Three gravid insects alight independently (for oviposition) at A, B , and C , where A, B , and C are uniformly distributed on the surface of the melon. For any two insects, if the distance between them (along a great circle of the melon) is less than $\pi/2$, then they detect each other's presence and will both fly off to seek an unoccupied melon. Show that the probability that exactly one insect is left in possession of the melon is $3(\pi - 1)/4\pi$, and that the probability that all three remain on the melon is $1/4\pi$.
- 15 Let X and Y have joint density $f(x, y) = c \sin(x + y), 0 < x, y < \pi/2$. Show that $c = \frac{1}{2}$; $\text{cov}(X, Y) = \frac{1}{2}(\pi - 2) - \frac{1}{16}\pi^2$, and $\rho(X, Y) = (8(\pi - 2) - \pi^2)/(\pi^2 + 8\pi - 32)$.
- 16 Let X and Y be independent exponential with parameters λ and μ , respectively. Now define $U = X \wedge Y$ and $V = X \vee Y$. Find $\mathbf{P}(U = X)$, and show that U and $V - U$ are independent.
- 17 Let $(X_i; i \geq 1)$ be independent with the uniform density on $(-1, 1)$. Let the density of $\sum_{i=1}^n X_i$ be $f_n(x)$. Show that

$$f_n(x) = \frac{1}{2} \int_{x-1}^{x+1} f_{n-1}(u) du \quad \text{for } n \geq 2,$$

and deduce that for any integer k , the density $f_n(x)$ is a polynomial in x for $x \in [k, k + 1)$.

- 18 Let X and Y have the bivariate normal density of Examples 8.4.3 and 8.20.
- (a) If $\sigma = \tau$, what is $\mathbf{E}(X|X + Y)$?
- (b) If $\sigma \neq \tau$, what are $\mathbf{E}(X|X + Y)$ and $\mathbf{E}(Y|X + Y)$?

19 Let $(X_n; n \geq 1)$ be independent and uniformly distributed on $(0, 1)$. Define

$$T = \min \left\{ n: \sum_{i=1}^n X_i > 1 \right\}.$$

Show that $\mathbf{P}(T \geq j + 1) = 1/j!$ for $j \geq 1$. Deduce that $\mathbf{E}(T) = e$.

20 Let (X_1, X_2, X_3) be independent and uniformly distributed on $(0, 1)$. What is the probability that the lengths X_1, X_2, X_3 can form a triangle?

21 Let $(U_i; i \geq 1)$ be independently and uniformly distributed on $(0, 1)$, and define $M_n = \max \{U_1, \dots, U_n\}$. Show that, as $n \rightarrow \infty$, the distribution of $Z_n = n(1 - M_n)$ converges to an exponential distribution.

22 Let $(X_i; i \geq 1)$ be independent exponential random variables each with parameter μ . Let N be independent of the X_i having mass function $f_N(n) = (1 - p)p^{n-1}; n \geq 1$. What is the density of $Y = \sum_{i=1}^N X_i$?

23 Let $N(t)$ be a Poisson process, $C(t)$ its current life at t , and X_1 the time of the first event. Show that $\text{cov}(X_1, C(t)) = \frac{1}{2}t^2 e^{-\lambda t}$.

24 **Simulating Gamma** Let U and X be independent, where U is uniform on $(0, 1)$ and X is exponential with parameter $\alpha^{-1} \leq 1$. Show that the density of X conditional on

$$\left(\frac{eX}{\alpha}\right)^{\alpha-1} \exp\left(-\left(\frac{\alpha-1}{\alpha}\right)X\right) \geq U$$

is $x^{\alpha-1}e^{-x}/\Gamma(\alpha)$. Why is this of value?

25 (a) Let X be exponential with parameter 1. Show that X/λ is exponential with parameter λ .

(b) Let $(X_i; 1 \leq i \leq n)$ be independent and exponential with parameter 1. Use the lack-of-memory property of the exponential density to show that $\max \{X_1, \dots, X_n\}$ has the same distribution as $X_1 + X_2/2 + \dots + X_n/n$.

26 Let X_1, X_2, X_3 , and X_4 be independent standard normal random variables. Show that $W = X_1/X_2$ has the Cauchy density, and $Z = |X_1 X_2 + X_3 X_4|$ has an exponential density.

27 Let X and Y be independent Poisson random variables each with parameter n . Show that, as $n \rightarrow \infty$, $\mathbf{P}(X - Y \leq \sqrt{2nx}) \rightarrow \Phi(x)$.

28 Let $(U_i; i \geq 1)$ be a collection of independent random variables each uniform on $(0, 1)$. Let X have mass function $f_X(x) = (e - 1)e^{-x}; x \geq 1$ and let Y have mass function $f_Y(y) = 1/\{(e - 1)y!\}; y \geq 1$. Show that $Z = X - \max \{U_1, \dots, U_Y\}$ is exponential. (Assume X and Y are independent of each other and of the U_i).

29 Let X and Y be independent gamma with parameters $(\alpha, 1)$ and $(\beta, 1)$, respectively. Find the conditional density of X given $X + Y = z$.

30 Let X and Y be independent standard normal random variables. Show that the pair X and Z , where $Z = \rho X + (1 - \rho^2)^{1/2} Y, |\rho| \leq 1$, has a standard bivariate normal density.

31 Let X and Y have joint moment generating function $M(s, t)$ and define $K(s, t) = \log M(s, t)$. Show that $K_s(0, 0) = \mathbf{E}(X)$, $K_{ss}(0, 0) = \text{var}(X)$, and $K_{st}(0, 0) = \text{cov}(X, Y)$.

32 A sorcerer has hidden a ring in one of an infinite number of boxes numbered $\dots, -2, -1, 0, 1, 2, 3, \dots$. You only have time to look in 11 boxes. The sorcerer gives you a hint. He tosses 100 fair coins and counts the number of heads. He does not tell you this number, nor does he tell you the number of the box with the ring in it, but he tells you the sum of these two numbers.

(a) If the sum is 75, which 11 boxes should you look in?

(b) Give an approximation to the probability of finding the ring.

$$\left(\frac{1}{\sqrt{2\pi}} \int_0^{11/10} e^{-\frac{u^2}{2}} du = 0.36.\right)$$

- 33 Multivariate Normal Density** Let (Y_1, \dots, Y_n) be independent, each having the $N(0, 1)$ density. If

$$X_i = \sum_j a_{ij} Y_j + b_i \quad \text{for } 1 \leq i, j \leq n,$$

then (X_1, \dots, X_n) are said to have a multivariate normal density. Find the joint m.g.f. $M_n(t_1, \dots, t_n) = \mathbf{E}(\exp(\sum_{i=1}^n t_i X_i))$. Deduce that the following three statements are equivalent:

- (a) The random variables (X_1, \dots, X_n) are independent.
 (b) (X_1, \dots, X_n) are pairwise independent.
 (c) $\text{cov}(X_i, X_j) = 0$ for $1 \leq i \neq j \leq n$.
- 34** A sequence of random variables X_1, X_2, \dots is said to obey the Central Limit Theorem (CLT) if and only if the distribution of $(S_n - \mathbf{E}(S_n))/\sqrt{\text{var}(S_n)}$ tends to the standard normal distribution, where $S_n = \sum_{i=1}^n X_i$.

State sufficient conditions on (X_n) for the sequence to obey the CLT and say which of your conditions are necessary.

Let $(U_n(\lambda_n))$ be a sequence of independent random variables having the Poisson distribution with nonzero means (λ_n) . In each of the following cases, determine whether the sequence (X_n) obeys the CLT:

- (i) $X_n = U_n(1)$. (iv) $X_n = U_{2n}(1)/(1 + U_{2n-1}(1))$.
 (ii) $X_n = U_n(1) + n$. (v) $X_n = U_n(n)$.
 (iii) $X_n = U_n(\frac{1}{2})!$.
- 35** Let X_1 and X_2 be independent with the same density $f(x)$. Let U be independent of both and uniformly distributed on $(0, 1)$. Let $Y = U(X_1 + X_2)$. Find $f(x)$ such that Y can also have density $f(x)$.

- 36 Molecules** A molecule M has velocity $v = (v_1, v_2, v_3)$ in Cartesian coordinates. Suppose that v_1, v_2 , and v_3 have joint density:

$$f(x, y, z) = (2\pi\sigma^2)^{-\frac{3}{2}} \exp\left(-\frac{1}{2\sigma^2}(x^2 + y^2 + z^2)\right).$$

Show that the density of the magnitude $|v|$ of v is

$$f(w) = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \sigma^{-3} w^2 \exp\left(-\frac{1}{2\sigma^2} w^2\right), \quad w > 0.$$

- 37** Let C be a circle radius r with centre O . Choose two points P and Q independently at random in C . Show that the probability that the triangle OPQ contains an obtuse angle is $\frac{3}{2}$. (Note: No integration is required.)
- 38** Given a fixed line AB , a point C is picked at random such that $\max\{AC, BC\} \leq AB$. Show that the probability that the triangle ABC contains an obtuse angle is

$$\frac{1}{8}\pi \left(\frac{1}{3}\pi - \frac{1}{4}\sqrt{3}\right)^{-1}.$$

[Note: No integration is required. This is a version of a problem given by Lewis Carroll. To combat insomnia, he solved mathematical problems in his head; this one was solved on the night of 20 January 1884. He collected a number of these mental exercises in a book entitled *Pillow problems* (Macmillan, 1895).]

- 39** Let X be a nonnegative random variable such that $\mathbf{P}(X > x) > 0$ for all $x > 0$. Show that $\mathbf{P}(X > x) \leq \mathbf{E}X^n/x^n$ for all $n \geq 0$. Deduce that $s = \sum_{n=0}^{\infty} \frac{1}{\mathbf{E}X^n} < \infty$.

Let N be a random variable with the mass function $\mathbf{P}(N = n) = \frac{1}{s\mathbf{E}X^n}$. Show that: (a) for all $x > 0$, $\mathbf{E}x^N < \infty$; (b) $\mathbf{E}X^N = \infty$.

- 40 Let X and Y be independent and uniform on $[0, 1]$, and let Z be the fractional part of $X + Y$. Show that Z is uniform on $[0, 1]$ and that X, Y, Z are pairwise independent but not independent.
- 41 Let $(X_{(k)}; 1 \leq k \leq n)$ be the order statistics derived from n independent random variables, each uniformly distributed on $[0, 1]$. Show that

$$(a) \quad \mathbf{E}X_{(k)} = \frac{k}{n+1}.$$

$$(b) \quad \text{var } X_{(k)} = \frac{k(n-k+1)}{(n+1)^2(n+2)}.$$

- 42 Let $(X_{(k)}; 1 \leq k \leq n)$ be the order statistics derived from n independent random variables each uniformly distributed on $[0, 1]$. Show that they have the same distribution as $(Y_k; 1 \leq k \leq n)$, where $Y_0 = 0$ and, given Y_{j-1}, Y_j , has the density $(1-y)^{n-j}; Y_{j-1} \leq y \leq 1$, for $1 \leq j \leq n$.
- 43 **Normal Sample:** Example 8.7.4 Revisited Let $(X_r; 1 \leq r \leq n)$ be independent $N(\mu, \sigma^2)$ random variables.

(a) By considering the joint moment generating function of \bar{X} and $(X_r - \bar{X}; 1 \leq r \leq n)$, show that $\bar{X} = \frac{1}{n} \sum_{r=1}^n X_r$ and

$$S^2 = \frac{1}{n-1} \sum_{r=1}^n (X_r - \bar{X})^2$$

are independent.

(b) Show that \bar{X} and $X_r - \bar{X}$ are uncorrelated, and deduce that \bar{X} and S^2 are independent.

- 44 (i) Suppose that the random variable Q has density $\sin q$, on $[0, \pi]$. Find the distribution function of Q , and deduce that $\sin^2 \{Q/2\}$ has the uniform density on $[0, 1]$.
- (ii) Suppose that the random vector (X, Y) is uniformly distributed on the unit circle, and set $R^2 = X^2 + Y^2$. Show that R^2 has the uniform density on $[0, 1]$. Deduce that the random vector (U, V, W) is uniformly distributed on the unit sphere, where $U = 2X(1 - R^2)^{\frac{1}{2}}$, $V = 2Y(1 - R^2)^{\frac{1}{2}}$, and $W = 1 - 2R^2$.

9

Markov Chains

In all crises of human affairs there are two broad courses open to a man. He can stay where he is or he can go elsewhere.

P.G. Wodehouse, Indiscretions of Archie

9.1 The Markov Property

In previous chapters, we found it useful and interesting to consider sequences of independent random variables. However, many observed sequences in the natural world are patently not independent. Consider, for example, the air temperature outside your window on successive days or the sequence of morning fixes of the price of gold. It is desirable and necessary to consider more general types of sequences of random variables.

After some thought, you may agree that for many such systems it is reasonable to suppose that, if we know exactly the state of the system today, then its state tomorrow should not further depend on its state yesterday (or on any previous state). This informal (and vague) preamble leads to the following formal (and precise) statement of the Markov property for a sequence of random variables.

(1) **Definition** Let $X = (X_n; n \geq 0)$ be a sequence of random variables taking values in a countable set S , called the state space. If for all $n \geq 0$ and all possible values of $i, k, k_0, \dots, k_{n-1}$, we have

(2)
$$\begin{aligned} \mathbf{P}(X_{n+1} = k | X_0 = k_0, \dots, X_n = i) &= \mathbf{P}(X_{n+1} = k | X_n = i) \\ &= \mathbf{P}(X_1 = k | X_0 = i), \end{aligned}$$

then X is said to be a Markov chain or to have the Markov property. We write

$$p_{ik} = \mathbf{P}(X_1 = k | X_0 = i),$$

where $(p_{ik}; i \in S, k \in S)$ are known as the transition probabilities of the chain. ▲

Sometimes we write $p_{i,k}$ for p_{ik} , and you are warned that some books use p_{ki} to denote p_{ik} .

Another popular rough and ready way of interpreting the formal condition (2) is to say that, for a Markov chain, the future is conditionally independent of the past, given the present.

Notice that in some applications it is more natural to start the clock at $n = 1$, so the chain is $X = (X_n; n \geq 1)$. Occasionally, it is convenient to suppose the chain extends in both directions so that $X = (X_n; -\infty < n < \infty)$.

The state space S is often a subset of the integers \mathbb{Z} or a subset of the set of ordered pairs of integers \mathbb{Z}^2 . Markov chains may take values in some countable set that happens not to be a subset of the integers. However, this set can immediately be placed in one–one correspondence with some appropriate subset of the integers, and the states relabelled accordingly.

- (3) **Example: Simple Random Walk** Let $(S_n; n \geq 0)$ be a simple random walk. Because the steps $(S_{n+1} - S_n; n \geq 0)$ are independent, the sequence S_n clearly has the Markov property, and the transition probabilities are given by

$$p_{ik} = \mathbf{P}(S_{n+1} = k | S_n = i) = \begin{cases} p & \text{if } k = i + 1 \\ q & \text{if } k = i - 1 \\ 0 & \text{otherwise.} \end{cases}$$

The state space S is the set of integers \mathbb{Z} . ●

- (4) **Example: Branching Process** Let Z_n be the size of the n th generation in an ordinary branching process. Because family sizes are independent, $Z = (Z_n; n \geq 0)$ is a Markov chain. The transition probabilities are given by

$$p_{ik} = \mathbf{P}(Z_{n+1} = k | Z_n = i) = \mathbf{P} \left[\sum_{r=1}^i Y_r = k \right],$$

where Y_1, \dots, Y_i are the i families of the n th generation given that $Z_n = i$. The state space is the set of nonnegative integers \mathbb{Z}^+ . ●

When S is a finite set, X is known as a finite Markov chain. Until further notice, we consider finite chains (unless it is specifically stated otherwise) and write $|S| = d$.

- (5) **Example: Information Source** A basic concern of telecommunications engineers is the transmission of signals along a channel. Signals arise at a *source*, and to devise efficient methods of communication it is necessary to have models for such sources. In general, it is supposed that the source produces a sequence of symbols randomly drawn from a finite alphabet A . By numbering the symbols from 1 to $|A|$, the output becomes a sequence of random variables $(X_n; n \geq 1)$ called a message. Various assumptions can be made about the output of sources, but a common and profitable assumption is that they have the Markov property. In this case, the output is a finite Markov chain and the source is called a simple Markov source. ●

Having formally defined a Markov chain X , we emphasize that there are many ways of presenting the idea of a Markov chain to the mind's eye. You should choose the one that

best suits the context of the problem and your own psyche. For example:

- (i) A particle performs a random walk on the vertices of a graph. The distribution of its next step depends on where it is, but not on how it got there.
- (ii) A system may be in any one of d states. The distribution of its next state depends on its current state, but not on its previous states.

Because of this imagery, we talk equivalently of chains visiting k , being at k , taking the value k , and so on. Whatever the choice of concept, the notation is always essentially that of Definition 1, but (to avoid repetitive strain injury) some abbreviations of notation are widespread. Thus, we commonly write

$$\mathbf{P}(X_{n+1} = k | X_0 = k_0, \dots, X_{n-1} = k_{n-1}, X_n = i) = \mathbf{P}(X_{n+1} = k | X_0, \dots, X_n).$$

If we want to stress or specify the initial value of X , then we write $\mathbf{P}(X_{n+1} = k | X_0 = k_0, \dots, X_n)$, and so on.

Note that the Markov property as defined in (1) is equivalent to each of the following properties, which it is occasionally convenient to take as definitive. First:

$$(6) \quad \mathbf{P}(X_{n+m} = k | X_0, \dots, X_n) = \mathbf{P}(X_{n+m} = k | X_n)$$

for any positive m and n . Second:

$$(7) \quad \mathbf{P}(X_{n_r} = k | X_{n_1}, \dots, X_{n_{r-1}}) = \mathbf{P}(X_{n_r} = k | X_{n_{r-1}})$$

for any $n_1 < n_2 < \dots < n_r$. Third:

$$(8) \quad \begin{aligned} \mathbf{P}(X_1 = k_1, \dots, X_{r-1} = k_{r-1}, X_{r+1} = k_{r+1}, \dots, X_n = k_n | X_r = k_r) \\ = \mathbf{P}(X_1 = k_1, \dots, X_{r-1} = k_{r-1} | X_r = k_r) \times \\ \times \mathbf{P}(X_{r+1} = k_{r+1}, \dots, X_n = k_n | X_r = k_r). \end{aligned}$$

You are asked to prove the equivalence of Definition 1 and (6), (7), and (8) in Problem 6. Notice that (8) expresses in a precise form our previously expressed rough idea that given the present state of a Markov chain, its future is independent of its past. Finally, it should be noted that the Markov property is preserved by some operations, but not by others, as the following examples show.

- (9) **Example: Sampling** Let X be a Markov chain. Show that the sequence $Y_n = X_{2n}$; $n \geq 0$, is a Markov chain.

Solution Because X is a Markov chain, we can argue as follows:

$$\begin{aligned} \mathbf{P}(Y_{n+1} = k | Y_0, \dots, Y_n = i) &= \mathbf{P}(X_{2n+2} = k | X_0, \dots, X_{2n} = i) \\ &= \mathbf{P}(X_{2n+2} = k | X_{2n} = i) \quad \text{by (7)} \\ &= \mathbf{P}(Y_{n+1} = k | Y_n = i). \end{aligned}$$

So Y is a Markov chain. It is said to be *imbedded* in X . ●

(10) **Example** If X is a Markov chain with state space S_X , show that

$$Y_n = (X_n, X_{n+1}); \quad n \geq 0,$$

is a Markov chain. What are its transition probabilities?

Solution The state space of Y is a collection of ordered pairs of the states of X ; that is to say, $S_Y = \{(s_1, s_2): s_1 \in S_X, s_2 \in S_X\}$. Now

$$\begin{aligned} \mathbf{P}(Y_{n+1} = (j, k) | Y_0, \dots, Y_n) &= \mathbf{P}(X_{n+2} = k, X_{n+1} = j | X_0, \dots, X_{n+1}) \\ &= \mathbf{P}(X_{n+2} = k, X_{n+1} = j | X_{n+1}, X_n) \quad \text{since } X \text{ is Markov,} \\ &= \mathbf{P}(Y_{n+1} = (j, k) | Y_n). \end{aligned}$$

So Y is Markov. Also,

$$(11) \quad \mathbf{P}(Y_{n+1} = (k, l) | Y_n = (i, j)) = \mathbf{P}(X_{n+2} = l | X_n = k) \delta_{jk} = p_{kl} \delta_{kj},$$

where

$$\delta_{kj} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}$$

is the usual Kronecker delta. ●

(12) **Example** Let X be a Markov chain. Show that $Y_n = |X_n|$; $n \geq 0$, is not necessarily a Markov chain.

Solution Let X have state space $S = \{-1, 0, 1\}$ and transition probabilities zero, except for

$$p_{-1,0} = \frac{1}{2}, p_{-1,1} = \frac{1}{2}, p_{0,-1} = 1, p_{1,0} = 1.$$

Then

$$\mathbf{P}(Y_{n+1} = 1 | Y_n = 1, Y_{n-1} = 1) = \mathbf{P}(X_{n+1} = 1 | X_n = 1, X_{n-1} = -1) = 0.$$

But $\mathbf{P}(Y_{n+1} = 1 | Y_n = 1) = \mathbf{P}(Y_{n+1} = 1 | X_n \in \{-1, 1\})$, which is not necessarily zero. So Y is not Markov. ●

Notice that the states -1 and $+1$ for X_n produce one state $+1$ for Y_n ; they are said to be *lumped* together. The example shows that lumping states together can destroy the Markov property. Conversely, given a sequence Y_n which is *not* a Markov chain, it is sometimes possible to construct a Markov chain involving Y_n by enlarging the state space.

(13) **Example** A machine can be in one of two states; working (denoted by s_0), or repair (denoted by s_1). Each day, if working, it may break down with probability α independently of other days. It takes r days to repair, where $r > 1$. Now if X_n is the state of the machine on the n th day, this is not a Markov chain. To see this note that

$$\mathbf{P}(X_{n+1} = s_0 | X_n = s_1, X_{n-1} = s_0) = 0,$$

but

$$\mathbf{P}(X_{n+1} = s_0 | X_n = X_{n-1} = \dots = X_{n-r+1} = s_1) = 1.$$

However, suppose we now let the state space be $S = \{s_0, s_1, \dots, s_r\}$, where $X_n = s_i$ if the machine has been in repair for i days. Then

$$\begin{aligned} \mathbf{P}(X_{n+1} = s_{i+1} | X_n = s_i, X_{n-1}, \dots) &= 1, \text{ if } 1 \leq i \leq r-1, \\ \mathbf{P}(X_{n+1} = s_0 | X_n = s_r, \dots) &= 1, \end{aligned}$$

and so on. It is easy to see that X_n now is a Markov chain. ●

9.2 Transition Probabilities

Recall that X is a Markov chain with state space S , where $|S| = d$. The transition probabilities p_{ik} are given by

$$(1) \quad p_{ik} = \mathbf{P}(X_{n+1} = k | X_n = i) \quad \text{for } n \geq 0.$$

The $d \times d$ matrix (p_{ij}) of transition probabilities is called the transition matrix and is denoted by \mathbf{P} . Let us first record two simple but important facts about \mathbf{P} . Because $(p_{ik}; k \in S)$ is a conditional mass function, we have

$$(2) \quad p_{ik} \geq 0 \quad \text{for all } i \text{ and } k;$$

and

$$(3) \quad \sum_{k \in S} p_{ik} = 1.$$

Any matrix \mathbf{P} satisfying (2) and (3) is called *stochastic*. We remark that if in addition

$$(4) \quad \sum_{i \in S} p_{ik} = 1,$$

then \mathbf{P} is *doubly stochastic*. Also, if (3) is replaced by the condition

$$(5) \quad \sum_{k \in S} p_{ik} \leq 1,$$

then a matrix satisfying (2) and (5) is called *substochastic*. For example, the simple random walk of Example 9.1.3 is doubly stochastic. If $p_{ij} > 0$ for all i and j , then \mathbf{P} is called *positive*.

Now, given that $X_0 = i$, the distribution of X_n is denoted by

$$p_{ik}(n) = \mathbf{P}(X_n = k | X_0 = i) = \mathbf{P}(X_{n+m} = k | X_m = i)$$

because of (9.1.2). Trivially, of course,

$$(6) \quad \sum_{k \in S} p_{ik}(n) = 1.$$

These probabilities are called the n -step transition probabilities, and they describe the random evolution of the chain.

Note that $p_{ij}(n)$ is a function of three variables, the two states i, j , and the time n . In more complicated expressions involving several such probabilities, you should use the symbols i, j, k, l to denote states and the symbols m, n, r, t to denote time (possibly with suffices).

Some simple special cases illustrate these notions.

- (7) **Example (9.1.3) Continued: Simple Random Walk** Recall that $(S_n; n \geq 0)$ are the successive values of a simple random walk. If $S_n = k$ and $S_0 = i$, then from Theorem 5.6.4 we have

$$(8) \quad p_{ik}(n) = \begin{cases} \binom{n}{\frac{1}{2}(n+k-i)} p^{\frac{1}{2}(n+k-i)} (1-p)^{\frac{1}{2}(n-k+i)} & \text{if } n+k-i \text{ is even} \\ 0 & \text{otherwise.} \end{cases}$$

(Note that this chain has infinite state space.) ■

- (9) **Example: Survival** A traffic sign stands in a vulnerable position. Each day, independently of other days, it may be demolished by a careless motorist with probability q . In this case, the city engineer replaces it with a new one at the end of the day. At the end of day n , let X_n denote the number of days since the sign in position was newly installed. Show that X_n is a Markov chain, and find p_{ik} and $p_{ik}(n)$.

(Note that this chain has infinite state space.)

Solution By construction,

$$X_{n+1} = \begin{cases} X_n + 1 & \text{with probability } 1 - q = p \\ 0 & \text{with probability } q. \end{cases}$$

Because the choice of outcomes is independent of previous days, X_n is a Markov chain, and

$$p_{ik} = \begin{cases} p & \text{if } k = i + 1 \\ q & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For the n -step transition probabilities, we note that either the sign survives for all n days or has been struck in the meantime. Hence,

$$(10) \quad p_{ik}(n) = p^n \quad \text{if } k = i + n$$

and

$$(11) \quad p_{ik}(n) = qp^k \quad \text{if } 0 \leq k \leq n - 1. \quad \bullet$$

Returning to the general case, we examine the relationship between p_{ik} and $p_{ik}(n)$. It is a remarkable and important consequence of the Markov property (9.1.1) that the random evolution of the chain is completely determined by p_{ik} , as the following theorem shows.

(12) **Theorem: Chapman–Kolmogorov Equations** *Let X have transition matrix P . For any i and k in S , and any positive m and n , we have*

$$(13) \quad p_{ik}(m+n) = \sum_{j \in S} p_{ij}(m)p_{jk}(n)$$

and also

$$(14) \quad p_{ik}(n+1) = \sum_{j_1 \in S} \cdots \sum_{j_n \in S} p_{ij_1} p_{j_1 j_2} \cdots p_{j_n k}.$$

Proof Recall that if $(A_j; j \leq d)$ is a collection of disjoint events such that $\cup_1^d A_j = \Omega$, then for any events B and C

$$\mathbf{P}(B|C) = \sum_{j=1}^d \mathbf{P}(B \cap A_j|C).$$

Hence, setting $A_j = \{X_m = j\}$, we have

$$\begin{aligned} p_{ik}(m+n) &= \sum_{j \in S} \mathbf{P}(X_{m+n} = k, X_m = j | X_0 = i) \\ &= \sum_{j \in S} \mathbf{P}(X_{m+n} = k | X_m = j, X_0 = i) \mathbf{P}(X_m = j | X_0 = i), \\ &\hspace{20em} \text{by conditional probability,} \\ &= \sum_{j \in S} \mathbf{P}(X_{m+n} = k | X_m = j) \mathbf{P}(X_m = j | X_0 = i), \\ &\hspace{20em} \text{by the Markov property,} \\ &= \sum_{j \in S} p_{ij}(m)p_{jk}(n). \end{aligned}$$

Hence, in particular,

$$(15) \quad \begin{aligned} p_{ik}(n+1) &= \sum_{j_1 \in S} p_{ij_1} p_{j_1 k}(n) \\ &= \sum_{j_1 \in S} \sum_{j_2 \in S} p_{ij_1} p_{j_1 j_2} p_{j_2 k}(n-1) \\ &\quad \vdots \\ &= \sum_{j_1 \in S} \cdots \sum_{j_n \in S} p_{ij_1} \cdots p_{j_n k} \end{aligned}$$

by repeated application of (15).

An alternative proof of (14) is provided by the observation that the summation on the right-hand side is the sum of the probabilities of all the distinct paths of n steps, which lead from i to k . Because these are mutually exclusive and one of them must be used to make the trip from i to k , the result follows. ■

The n -step transition probabilities $p_{ik}(n)$ tell us how the mass function of X_n depends on X_0 . If X_0 itself has mass function

$$(16) \quad \alpha_i = \mathbf{P}(X_0 = i)$$

and X_n has mass function

$$(17) \quad \alpha_i^{(n)} = \mathbf{P}(X_n = i)$$

then, by conditional probability, they are related by

$$(18) \quad \alpha_k^{(n)} = \sum_{j \in S} \alpha_j p_{jk}(n).$$

The probabilities $\alpha_i^{(n)}$ are sometimes called the *absolute probabilities* of X_n .

Now we notice that the d^2 n -step transition probabilities ($p_{ik}(n)$; $1 \leq i \leq d$, $1 \leq k \leq d$) can be regarded as a matrix \mathbf{P}_n , and the absolute probabilities ($\alpha_i^{(n)}$; $1 \leq i \leq d$) as a row vector α_n . It follows from Theorem 12 and (18) that

$$(19) \quad \mathbf{P}_{m+n} = \mathbf{P}_m \mathbf{P}_n = \mathbf{P}^{m+n}$$

and $\alpha_n = \alpha \mathbf{P}^n$, where $\alpha = (\alpha_1, \dots, \alpha_d)$.

(20) Example: Two State Chain The following simple but important example is very helpful in illustrating these and other ideas about Markov chains. Let X have state space $S = \{1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}.$$

You can verify by induction that

$$(\alpha + \beta) \mathbf{P}^n = \begin{pmatrix} \beta & \alpha \\ \beta & \alpha \end{pmatrix} + (1 - \alpha - \beta)^n \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix}.$$

Hence, for example,

$$p_{12}(n) = \frac{\alpha}{\alpha + \beta} - \frac{\alpha}{\alpha + \beta} (1 - \alpha - \beta)^n. \quad \bullet$$

Descending once again from the general to the particular, we identify some special varieties of chain that have attractive properties that we find useful later.

(21) Definition *If for some $n_0 < \infty$ we have $p_{ij}(n_0) > 0$ for all i and j , then the chain is said to be regular.* ▲

(22) Example Let X have transition probabilities

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Then \mathbf{P} is not positive, but

$$\mathbf{P}^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} > 0,$$

so \mathbf{P} is regular. ●

Roughly speaking, a chain is regular if there is a time such that, no matter where it started, the chain could be anywhere in S . Some chains satisfy the weaker condition that every state can be reached from every other state with nonzero probability. This is called *irreducibility*.

(23) Definition A chain X is irreducible if for each i and k in S there exists an $n_0 < \infty$, such that $p_{ik}(n_0) > 0$. ▲

(24) Example Let X have transition matrix

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then,

$$\mathbf{P}^{2n} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad n \geq 0$$

and

$$\mathbf{P}^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \quad n \geq 0.$$

Hence, X is neither positive nor regular but it is irreducible. In fact, it is said to be *periodic* with period 2 because $p_{ii}(n) > 0$ if n is even. A state with no period greater than 1 is *aperiodic*. ●

(25) Example Let X and Y be independent regular Markov chains with transition matrices $\mathbf{P} = (p_{ik})$ and $\mathbf{Q} = (q_{jl})$, respectively. Show that $Z_n = (X_n, Y_n)$; $n \geq 0$, is a regular Markov chain.

Solution Using the independence of X and Y ,

$$\begin{aligned} \mathbf{P}(Z_n = (k, l) | Z_{n-1} = (i, j), Z_{n-2}, \dots, Z_0) &= \mathbf{P}(X_n = k | X_{n-1} = i, X_{n-2}, \dots, X_0) \\ &\quad \times \mathbf{P}(Y_n = l | Y_{n-1} = j, \dots, Y_0) \\ &= p_{ik} q_{jl} \end{aligned}$$

because X and Y are Markov chains. Therefore, Z is a Markov chain. Likewise, Z has n -step transition probabilities $p_{ik}(n)q_{jl}(n)$.

Finally, because \mathbf{P} and \mathbf{Q} are regular, there exists n_0 and m_0 (both finite) such that $p_{ik}(m_0)$ and $q_{jl}(n_0)$ are both positive for all i, k and all j, l , respectively. Hence, $p_{ik}(m_0 n_0) q_{jl}(m_0 n_0) > 0$, for all i, j, k, l , and so Z is regular. ●

Note two further bits of jargon. A set C of states is called *closed* if $p_{ik} = 0$ for all $i \in C, k \notin C$. Furthermore, if C is closed and $|C| = 1$, then this state is called *absorbing*.

We conclude this section with two examples drawn from communication theory.

- (26) **Example: Entropy** Let the random vector $X_n = (X_0, \dots, X_n)$ have joint mass function $f(x_0, \dots, x_n)$. Then the *entropy* (also called *uncertainty*) of X_n is defined as

$$H(X_n) = -\mathbf{E}[\log(f(X_0, \dots, X_n))]$$

(with the convention that $0 \log 0 = 0$). Let X_0, \dots, X_n be the first $n + 1$ values of a Markov chain with transition matrix \mathbf{P} and initial mass function α . Show that, in this case,

$$(27) \quad H(X_n) = -\mathbf{E}[\log(\alpha_{X_0})] - \sum_{r=1}^n \mathbf{E}[\log(p_{X_{r-1}X_r})].$$

Solution Because X is a Markov chain, $f(x_0, x_1, \dots, x_n) = \alpha_{x_0} p_{x_0x_1} \dots p_{x_{n-1}x_n}$. Hence,

$$\begin{aligned} & \mathbf{E}[\log(f(X_0, \dots, X_n))] \\ &= \sum_{x_0 \in S} \dots \sum_{x_n \in S} \alpha_{x_0} p_{x_0x_1} \dots p_{x_{n-1}x_n} (\log \alpha_{x_0} + \log p_{x_0x_1} + \dots + \log p_{x_{n-1}x_n}) \\ &= \sum_{x_0} \alpha_{x_0} \log \alpha_{x_0} + \sum_{x_0, x_1} \alpha_{x_0} p_{x_0x_1} \log p_{x_0x_1} + \dots + \sum_{x_{n-1}, x_n} \alpha_{x_{n-1}} p_{x_{n-1}x_n} \log p_{x_{n-1}x_n} \\ &= \mathbf{E}[\log(\alpha_{X_0})] + \sum_{r=1}^n \mathbf{E}[\log p_{X_{r-1}X_r}], \end{aligned}$$

as required, yielding (27). ●

- (28) **Example: Simple Markov Source** Let the random variable X and the random vector Y be jointly distributed, and denote the conditional mass function of X given Y by $f(x|y)$. Then the *conditional entropy* of X with respect to Y is defined to be

$$\begin{aligned} H(X|Y) &= -\mathbf{E}[\mathbf{E}(\log f(X|Y)|Y)] \\ &= -\sum_y \sum_x f(x|y) \log f(x|y) \mathbf{P}(Y = y). \end{aligned}$$

Let X_0, \dots, X_{n+1} be the output from the Markov source defined in Example 9.1.5. Show that

$$H(X_{n+1}|X_0, \dots, X_n) = H(X_{n+1}|X_n).$$

Solution Let Y be (X_0, \dots, X_n) . Then, by the Markov property,

$$\begin{aligned} f(x|y) &= \mathbf{P}(X_{n+1} = x | X_0 = y_0, \dots, X_n = y_n) = \mathbf{P}(X_{n+1} = x | X_n = y_n) \\ &= p_{y_n x}. \end{aligned}$$

Hence,

$$\begin{aligned} H(X_{n+1}|Y) &= - \sum_y \sum_x p_{y_n x} \log p_{y_n x} \mathbf{P}(X_0 = y_0, \dots, X_n = y_n) \\ &= - \sum_{y_n} \sum_x p_{y_n x} \log p_{y_n x} \mathbf{P}(X_n = y_n) = \mathbf{H}(X_{n+1}|X_n). \quad \bullet \end{aligned}$$

9.3 First Passage Times

For any two states i and k of X , we are often interested in the time it takes for the chain to travel from i to k . This is not merely a natural interest, these quantities are also of theoretical and practical importance. For example, in the simple gambler's ruin problem the state 0 entails ruin, and in the simple branching process $X = 0$ entails extinction.

(1) **Definition** For a Markov chain X with $X_0 = i$:

(a) When $i \neq k$, the first passage time to k from i is defined to be

$$T_{ik} = \min \{n \geq 0 : X_n = k | X_0 = i\};$$

the mean first passage time is

(2)
$$\mu_{ik} = \mathbf{E}(T_{ik}).$$

(b) When $i = k$, the recurrence time of i is defined to be

$$T_i = \min \{n > 0 : X_n = i | X_0 = i\};$$

the mean recurrence time is

(3)
$$\mu_i = \mathbf{E}(T_i). \quad \blacktriangle$$

Note the simple but important fact that the chain has not entered k by time n if and only if

(4)
$$T_{ik} > n.$$

(5) **Example** Let X have transition matrix

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Then, given $X_0 = 1$, the chain enters 2 as soon as it leaves 1. Hence,

$$\mathbf{P}(T_{12} = r) = \frac{2}{3} \cdot \left(\frac{1}{3}\right)^{r-1}; \quad r \geq 1,$$

and

$$\mu_{12} = \sum_{r=1}^{\infty} r \frac{2}{3} \left(\frac{1}{3}\right)^{r-1} = \frac{3}{2}.$$

Likewise, first return to 1 at the r th step occurs after $r - 2$ consecutive visits to 2. so

$$\mathbf{P}(T_1 = r) = \begin{cases} \frac{1}{3} & ; r = 1 \\ \frac{2}{3} \left(\frac{3}{4}\right)^{r-2} \frac{1}{4} & ; r \geq 2. \end{cases}$$

Hence,

$$\mu_1 = \frac{1}{3} + \sum_{r=2}^{\infty} r \frac{1}{6} \left(\frac{3}{4}\right)^{r-2} = \frac{11}{3}.$$

If we do not require the distribution of T_{12} or T_1 , then a simpler procedure will suffice to find μ_{12} and μ_1 , as follows. Conditioning on the first step of the chain and assuming all the expectations exist, we find that

$$\mu_{12} = \frac{1}{3}\mathbf{E}(T_{12}|X_1 = 1) + \frac{2}{3}\mathbf{E}(T_{12}|X_1 = 2).$$

But, by the Markov property,

$$\mathbf{E}(T_{12}|X_1 = 1) = 1 + \mathbf{E}(T_{12}),$$

and obviously $\mathbf{E}(T_{12}|X_1 = 2) = 1$. Hence, $\mu_{12} = 1 + \frac{1}{3}\mu_{12}$ as above. Likewise, we find $\mu_{21} = 4$, and using conditional expectation again yields $\mu_1 = 1 + \frac{2}{3}\mu_{21} = \frac{11}{3}$ as before. ●

For a rather different type of behaviour consider the following.

(6) **Example** Let X have transition matrix

$$P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Because $p_{33} = 1$, state 3 is absorbing, which is to say that upon entering 3 the chain never leaves it subsequently. Hence, $T_{12} = r$ occurs when the first $r - 1$ visits to 1 are followed by a step to 2. Thus,

$$\mathbf{P}(T_{12} = r) = \left(\frac{1}{3}\right)^r ; \quad r \geq 1.$$

Hence,

$$\mathbf{P}(T_{12} < \infty) = \sum_{r=1}^{\infty} \left(\frac{1}{3}\right)^r = \frac{2}{3}$$

and $\mu_{12} = \infty$. Likewise,

$$\mathbf{P}(T_1 = r) = \begin{cases} \frac{1}{3} & ; r = 1 \\ \frac{1}{3} \left(\frac{3}{4}\right)^{r-2} \frac{1}{4} & ; r \geq 2, \end{cases}$$

and so $\mathbf{P}(T_1 < \infty) = \frac{2}{3}$ and $\mu_1 = \infty$. ●

These examples demonstrate that the properties of recurrence and first passage times depend strongly on the nature of the transition matrix \mathbf{P} . In fact, we are going to show that, for any finite regular chain, both μ_k and μ_{ik} are finite (with finite expectation) for all i and k . First we need to clear the ground a little. Because we are only considering finite chains with $|S| = d$, we can without loss of generality set $k = d$. (If you like mnemonics you can think of d as the destination of the chain.) Also, as we are only interested in the progress of the chain until it arrives at d , it is natural to focus attention on the probabilities

$$(7) \quad r_{ik}(n) = \mathbf{P}(X_n = k, n < T_{id} | X_0 = i), \quad i \neq d \neq k.$$

These are the transition probabilities of the chain before entering d , and we denote the array $(r_{ik}(n))$ by \mathbf{R}_n .

By definition, for one step,

$$(8) \quad r_{ik}(1) = p_{ik} \quad \text{for } i \neq d \neq k.$$

For $n > 1$, the n -step d -avoiding probabilities are given by the following.

(9) Theorem For $i \neq d \neq k$,

$$(10) \quad r_{ik}(n) = \sum_{j_1 \neq d} \sum_{j_2 \neq d} \cdots \sum_{j_{n-1} \neq d} p_{ij_1} p_{j_1 j_2} \cdots p_{j_{n-1} k},$$

or in matrix form $\mathbf{R}_n = \mathbf{R}_1^n$.

Proof We use the idea of paths. Every distinct path of the chain that goes from i to k in n steps and does not enter d is of the form $i, j_1, j_2, \dots, j_{n-1}, k$, where $j_r \in S \setminus d$ for $1 \leq r \leq n-1$. Such a path has probability $p_{ij_1} \cdots p_{j_{n-1} k}$ and one of them is used, so $r_{ik}(n)$ is just the sum of all these probabilities as given on the right-hand side of (10). ■

Corollary For any state i of a regular chain,

$$(11) \quad \lim_{n \rightarrow \infty} \sum_{k \neq d} r_{ik}(n) = 0,$$

and more strongly,

$$(12) \quad \sum_{n=1}^{\infty} \sum_{k \neq d} r_{ik}(n) < \infty.$$

Proof First, suppose that the chain is positive so that for every i , $p_{id} > 0$. Hence, there exists t such that

$$(13) \quad \sum_{k \neq d} p_{ik} < t < 1.$$

Therefore, using (13) on the last sum in (10),

$$\sum_{k \neq d} r_{ik}(n) \leq \sum_{j_1 \neq d} \cdots \sum_{j_{n-1} \neq d} p_{ij_1} \cdots p_{j_{n-2} j_{n-1} t}.$$

Hence,

$$\sum_{k \neq d} r_{ik}^{(n)} \leq t^n$$

on using (13) to bound each summation successively. Because $t < 1$, (11) and (12) follow in this case. If the chain is regular but not positive, we first note that because $\sum_k r_{jk} \leq 1$ we have that

$$\sum_{k \neq d} r_{ik}(n+1) = \sum_{j \neq d} r_{ij}(n) \sum_k r_{jk} \leq \sum_{j \neq d} r_{ij}(n).$$

Thus, $\sum_k r_{ik}(n)$ is nonincreasing in n . Because the chain is regular, there is an m_0 such that $p_{id}(m_0) > 0$ for all i . By the argument of the first part, for some $t_0 < 1$,

$$(14) \quad \sum_{k \neq d} r_{ik}(nm_0) < t_0^n < 1.$$

Hence, because $\sum_k r_{ik}(n)$ is nondecreasing (11) follows. Finally,

$$\sum_n \sum_k r_{ik}(n) \leq m_0 \left(1 + \sum_{n,k} r_{ik}(m_0 n) \right) \leq m_0 \left(\frac{1}{1-t_0} \right) < \infty$$

proving (12). ■

With these preliminaries completed, we can get on with proving the main claim of the paragraph following Example 6.

(15) **Theorem** *For a regular chain, T_{id} is finite with probability 1 and has finite mean. More precisely, $\mathbf{P}(T_{id} > n) < c\lambda^n$ for some constants $c < \infty$ and $\lambda < 1$.*

Proof By the remark preceding (4),

$$\mathbf{P}(T_{id} > n) = \sum_{k \neq d} r_{ik}(n) \rightarrow 0$$

as $n \rightarrow \infty$ by (11). Therefore, T_{id} is finite with probability 1. Also,

$$\mathbf{E}(T_{id}) = \sum_{n=0}^{\infty} \mathbf{P}(T_{id} > n) = \sum_{n=0}^{\infty} \sum_{k \neq d} r_{ik}(n) < \infty$$

by (12). The second statement of the theorem follows easily from (14). ■

There is a simple and useful generalization of this result, as follows.

(16) **Theorem** *Let X be a regular Markov chain, and let D be a subset of the state space S . For $i \notin D$, define the first passage time $T_{iD} = \min \{n: X_n \in D | X_0 = i\}$. Then, $\mathbf{E}(T_{iD}) < \infty$.*

Proof This is an exercise for you. ■

It should be remarked that $\mathbf{E}(T_i) < \infty$ is a trivial consequence of Theorem 15.

As discussed above, first passage times are interesting in themselves for practical reasons, but they are even more interesting because of a crucial theoretical property. Informally, it says that given the state of a chain at a first passage time T , the future of the chain is independent of the past. The following example makes this more precise.

(17) Example: Preservation of Markov Property at First Passage Times Let X be a regular Markov chain with transition matrix \mathbf{P} , and let T be the first passage time of the chain to d . Show that for any $m > 0$ and $x_r \neq d$, we have

$$(18) \quad \mathbf{P}(X_{T+m} = k | X_r = x_r \text{ for } 1 \leq r \leq T, X_T = d) = p_{dk}(m).$$

Solution Let us denote the event $\{X_r = x_r \neq d \text{ for } 1 \leq r < T\}$ by $A(T)$. Then, using conditional probability, the left-hand side of (18) may be written as

$$(19) \quad \frac{\mathbf{P}(X_{T+m} = k, A(T), X_T = d)}{\mathbf{P}(A(T), X_T = d)}$$

Now the numerator can be expanded as

$$\begin{aligned} & \sum_{t=1}^{\infty} \mathbf{P}(X_{T+m} = k, A(t), X_t = d, T = t) \\ &= \sum_{t=1}^{\infty} \mathbf{P}(X_{t+m} = k | A(t), X_t = d) \mathbf{P}(A(t), X_t = d, T = t) \\ &= p_{dk}(m) \sum_{t=1}^{\infty} \mathbf{P}(A(t), X_t = d, T = t) \quad \text{by the Markov property,} \\ &= p_{dk}(m) \mathbf{P}(A(T), X_T = d). \end{aligned}$$

Finally, substitution into (19) yields (18). ●

It would be difficult to overemphasize the importance of this result in the theory of Markov chains; it is used repeatedly. [It is a special case of the “strong Markov property,” that we meet later.]

To conclude this section we show that the mass functions of T_{id} and of T_d are related to the transition probabilities $p_{ik}(n)$ by very elegant and useful identities. Let $f_{id}(n) = \mathbf{P}(T_{id} = n)$, $i \neq d$, and $f_{dd}(n) = \mathbf{P}(T_d = n)$. Define the generating functions

$$P_{ik}(z) = \sum_{n=0}^{\infty} p_{ik}(n)z^n \quad \text{and} \quad F_{id}(z) = \sum_{n=0}^{\infty} f_{id}(n)z^n$$

with the convention that $p_{ii}(0) = 1$, $p_{ij}(0) = 0$, for $i \neq j$, and $f_{ij}(0) = 0$ for all i and j .

(20) Theorem When $i \neq k$, we have

$$(21) \quad P_{ik}(z) = F_{ik}(z)P_{kk}(z),$$

and otherwise

$$(22) \quad P_{ii}(z) = 1 + F_{ii}(z)P_{ii}(z).$$

Proof The idea of the proof is much the same as that of Example 17. For each k in S , let us define the event $A_m = \{X_m = k\}$, and let B_m be the event that the first visit to k after time 0 takes place at time m . That is,

$$B_m = \{X_r \neq k \text{ for } 1 \leq r < m, X_m = k\}.$$

Then following a now familiar route, we write

$$(23) \quad \begin{aligned} p_{ik}(m) &= \mathbf{P}(A_m | X_0 = i) = \sum_{r=1}^m \mathbf{P}(A_m \cap B_r | X_0 = i) \\ &= \sum_{r=1}^m \mathbf{P}(A_m | B_r, X_0 = i) \mathbf{P}(B_r | X_0 = i) \\ &= \sum_{r=1}^m \mathbf{P}(A_m | X_r = k) \mathbf{P}(B_r | X_0 = i) \quad \text{by the Markov property,} \\ &= \sum_{r=1}^m p_{kk}(m-r) f_{ik}(r). \end{aligned}$$

The right-hand side of (23) is a convolution, so multiplying both sides by z^m and summing over all $m \geq 1$ gives $P_{ik}(z) - \delta_{ik} = F_{ik}(z)P_{kk}(z)$, where $\delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$ as required. ■

(24) **Example: Weather** Successive days are either hot or cold, and they are also either wet or dry. From one day to the next, either the temperature changes with probability α or the precipitation changes with probability $1 - \alpha$. Let $f(n)$ be the probability that it is again hot and dry for the first time on the n th day, given that it was hot and dry on day zero. Show that

$$(25) \quad F(z) = \sum_{n=1}^{\infty} f(n)z^n = z^2 \frac{1 + (1 - 2z^2)(1 - 2\alpha)^2}{2 - z^2 - z^2(1 - 2\alpha)^2}.$$

Solution It is helpful to visualize this Markov chain as a random walk on the vertices of a square in which steps are taken along a horizontal edge with probability α or a vertical edge with probability $1 - \alpha$. We identify the four states of the chain with the vertices of the square; the origin is hot and dry. The walk can return to the origin only after an even number $2n$ of steps, of which $2k$ are horizontal and $2n - 2k$ are vertical. Hence $p_0(2n)$, the probability of returning on the $2n$ th step (not necessarily for the first time), is

$$\begin{aligned} p_0(2n) &= \sum_{k=0}^n \alpha^{2k} (1 - \alpha)^{2n-2k} \binom{2n}{2k} = \frac{1}{2}(\alpha + (1 - \alpha))^{2n} + \frac{1}{2}(\alpha - (1 - \alpha))^{2n} \\ &= \frac{1}{2}((1 - 2\alpha)^{2n} + 1). \end{aligned}$$

Hence,

$$P_0(z) = \sum_0^{\infty} p_0(2n)z^{2n} = \frac{1}{2} \left(\frac{1}{1 - (1 - 2\alpha)^2 z^2} + \frac{1}{1 - z^2} \right).$$

Hence, by (22), we have

$$F(z) = \frac{P_0(z) - 1}{P_0(z)} = z^2 \frac{1 + (1 - 2z^2)(1 - 2\alpha)^2}{2 - z^2(1 + (1 - 2\alpha)^2)},$$

which is (25). ●

If you have read Section 6.7, you will have noticed much in common with the above analysis and the results of that section. This is, of course, because the visits of a Markov chain to some given state k form a renewal process. We explore this link a little in Example 9.14.

9.4 Stationary Distributions

We now consider one of the most important properties of the transition matrix \mathbf{P} . That is, for any $d \times d$ stochastic matrix \mathbf{P} , the set of equations

$$(1) \quad x_k = \sum_{1 \leq i \leq d} x_i p_{ik}; \quad 1 \leq k \leq d,$$

always has a solution such that

$$(2) \quad x_i \geq 0$$

and

$$(3) \quad \sum_{i=1}^d x_i = 1.$$

Such a solution is thus a probability mass function, and it is commonly denoted by $x = \pi = (\pi_1, \dots, \pi_d)$. It may not be unique.

(4) **Example** (a) If

$$\mathbf{P} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix},$$

then clearly $\pi = (\frac{1}{2}, \frac{1}{2})$.

(b) If

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then it is also clear that $\pi = (\frac{1}{2}, \frac{1}{2})$.

(c) If

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

then we have $\pi = (\alpha, 1 - \alpha)$ for any $\alpha \in [0, 1]$.

Note that the first chain is regular, the second periodic, and the third has two absorbing states; these chains evolve in very different ways. ●

The mass function π is called a *stationary distribution* of the chain for the following reason. Suppose that π is the mass function of X_0 , then X_1 has mass function

$$\alpha_k(1) = \sum_i \pi_i p_{ik} = \pi_k$$

because π is a solution of (1). Hence, X_1 has mass function π , and by a trivial induction so does X_n for all n :

$$(5) \quad \mathbf{P}(X_n = k) = \pi_k; \quad n \geq 0.$$

Remark The chain is sometimes said to be in *equilibrium*.

In formal terms, (1) says that P has a positive left eigenvector corresponding to the eigenvalue 1. Experience of student calculations leads us to stress that π is a *left* eigenvector. If your stationary vector π is constant, check that you have not inadvertently found the right eigenvector. [And see (22) below.]

Here is a less trivial example.

(6) **Example: Random Walk with Retaining Barriers** Let X have state space $\{0, 1, 2, \dots, d\}$ and transition probabilities

$$\begin{aligned} p_{i,i+1} &= p; & 0 \leq i \leq d-1 \\ p_{i,i-1} &= 1-p = q; & 1 \leq i \leq d \\ p_{00} &= q \\ p_{dd} &= p. \end{aligned}$$

Then a stationary distribution must satisfy

$$(7) \quad \begin{aligned} \pi_i &= p\pi_{i-1} + q\pi_{i+1}; & 1 \leq i \leq d-1 \\ p\pi_0 &= q\pi_1 \\ q\pi_d &= p\pi_{d-1}. \end{aligned}$$

Simple substitution shows that if $p \neq q$, then $\pi_i = \pi_0 \left(\frac{p}{q}\right)^i$. Because $\sum_i \pi_i = 1$, it now follows that

$$\pi_i = \frac{1 - \frac{p}{q}}{1 - \left(\frac{p}{q}\right)^{d+1}} \left(\frac{p}{q}\right)^i. \quad \bullet$$

(8) **Example** Let X have transition matrix \mathbf{P} , and suppose that there exists a stationary distribution π satisfying (1). Define the Markov chain Y_n by $Y_n = (X_n, X_{n+1})$; $n \geq 0$. Show that Y has stationary distribution

$$(9) \quad \eta_{ij} = \pi_i p_{ij}; \quad i \in S, j \in S.$$

Solution We just have to check that η satisfies (1) and (3). Recall from Example 9.1.10 that Y has transition probabilities

$$\mathbf{P}(Y_{n+1} = (k, l) | Y_n = (i, j)) = p_{kl} \delta_{jk},$$

so that

$$\sum_{i,j} \eta_{ij} p_{kl} \delta_{jk} = \sum_{i,j} \pi_i p_{ij} p_{kl} \delta_{jk} = \sum_j \pi_j p_{kl} \delta_{jk} = \pi_k p_{kl} = \eta_{kl}.$$

Furthermore,

$$\begin{aligned} \sum_{i,j} \eta_{ij} &= \sum_{i,j} \pi_i p_{ij} = \sum_j \pi_j \quad \text{by (1)} \\ &= 1 \quad \text{by (3)}. \end{aligned}$$

Hence, η is the stationary distribution of Y . ●

(10) **Example: Nonhomogeneous Random Walk** Let $(S_n; n \geq 0)$ be a Markov chain with transition matrix given by

$$\begin{aligned} p_{i,i+1} &= \lambda_i, \\ p_{i,i-1} &= \mu_i, \\ p_{i,k} &= 0, \quad \text{if } |i - k| \neq 1, \end{aligned}$$

where $\lambda_i + \mu_i = 1$. This may be regarded as a random walk, taking positive or negative unit steps on the integers, such that the step probabilities depend on the position of the particle.

Is there a stationary distribution π ?

For simplicity, let us suppose that $\mu_0 = 0$ and $S_0 \geq 0$, so that the walk is confined to the nonnegative integers. Then if π exists, it satisfies

$$\begin{aligned} \pi_0 &= \mu_1 \pi_1 \\ \pi_1 &= \lambda_0 \pi_0 + \mu_2 \pi_2 \\ \pi_2 &= \lambda_1 \pi_1 + \mu_3 \pi_3 \end{aligned}$$

and in general, for $k > 1$,

$$\pi_k = \lambda_{k-1} \pi_{k-1} + \mu_{k+1} \pi_{k+1}.$$

Solving these equations in order of appearance gives

$$\pi_1 = \frac{\lambda_0}{\mu_1} \pi_0; \pi_2 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} \pi_0; \pi_3 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} \pi_0;$$

and so on. It is now easy to verify that for $n > 0$,

$$\pi_n = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} \pi_0.$$

This is a stationary distribution if

$$1 = \sum_{n=0}^{\infty} \pi_n = \pi_0 + \pi_0 \sum_{n=1}^{\infty} \frac{\prod_{r=1}^{n-1} \lambda_r}{\prod_{r=1}^n \mu_r},$$

and so we deduce that a stationary distribution exists if this sum converges. ●

Having examined some consequences of (1), we now turn to the question of proving it.

The existence of x satisfying (1), (2), and (3) is a famous result with many algebraic and analytical proofs. Most of these are neither elementary nor probabilistic. We prefer to give a proof that uses the ideas of probability theory and is elementary.

(11) **Theorem** *A regular Markov chain with transition matrix \mathbf{P} has a stationary distribution π .*

Proof Let s be an arbitrary state of the chain with recurrence time T_s and mean recurrence time μ_s . For all $k \in S$, let $\rho_k(s)$ be the expected number of visits to k between successive visits to s ; with the convention that $\rho_s(s) = 1$. We show that

$$(12) \quad \pi_k = \mu_s^{-1} \rho_k(s); \quad 1 \leq k \leq d$$

is a stationary distribution of the chain.

First, let I_n denote the indicator of the event that the chain visits k at the n th step and has not previously revisited s , given that it started in s . Then the total number of visits to k between visits to s is

$$R_k = \sum_{n=1}^{\infty} I_n; \quad k \neq s,$$

and in accord with our convention above, when $k = s$, we have $R_s = 1$. Now

$$T_s = 1 + \sum_{k \neq s} R_k.$$

It follows that the expected value $\rho_k(s)$ of R_k is finite, and also that

$$(13) \quad \mu_s = \sum_k \rho_k(s).$$

Furthermore,

$$\rho_k(s) = \mathbf{E}(R_k) = \sum_{n=1}^{\infty} \mathbf{E}(I_n) = \sum_{n=1}^{\infty} \mathbf{P}(X_n = k, T_s \geq n | X_0 = s).$$

Now for $n = 1$,

$$\mathbf{P}(X_1 = k, T_s \geq 1 | X_0 = s) = p_{sk}.$$

For $n \geq 2$,

$$\begin{aligned} \mathbf{P}(X_n = k, T_s \geq n | X_0 = s) &= \sum_{j \neq s} \mathbf{P}(X_n = k, X_{n-1} = j, T_s \geq n | X_0 = s) \\ &= \sum_{j \neq s} \mathbf{P}(X_n = k | X_{n-1} = j, T_s \geq n, X_0 = s) \mathbf{P}(X_{n-1} = j, T_s \geq n | X_0 = s) \\ &\hspace{15em} \text{by conditional probability} \\ &= \sum_{j \neq s} p_{jk} \mathbf{P}(X_{n-1} = j, T_s \geq n - 1 | X_0 = s) \quad \text{by the Markov property.} \end{aligned}$$

Hence,

$$\begin{aligned} (14) \quad \rho_k(s) &= p_{sk} + \sum_{j \neq s} p_{jk} \sum_{n=2}^{\infty} \mathbf{P}(X_{n-1} = j, T_s \geq n - 1 | X_0 = s) \\ &= \rho_s(s) p_{sk} + \sum_{j \neq s} p_{jk} \rho_j(s) = \sum_j \rho_j(s) p_{jk}. \end{aligned}$$

Dividing throughout by μ_s yields the result (12), as required. ■

In view of the appearance of mean recurrence times in the above proof, it is perhaps not surprising to discover another intimate link between π and μ .

(15) **Theorem** *For a regular Markov chain, the stationary distribution is unique and satisfies $\pi_k \mu_k = 1; k \in S$. Hence,*

$$(16) \quad \rho_k(s) = \frac{\mu_s}{\mu_k}.$$

Proof Recall that $T_{ik} = \min \{n \geq 0: X_n = k | X_0 = i\}$, so that, in particular, $T_{kk} = 0$ and $T_k = \min \{n \geq 1: X_n = k | X_0 = k\}$. Conditioning on the outcome of the first transition of the chain we have, for $i \neq k$,

$$(17) \quad \mu_{ik} = \mathbf{E}(\mathbf{E}(T_{ik} | X_1)) = 1 + \sum_j p_{ij} \mu_{jk}.$$

Also,

$$(18) \quad \mu_k = 1 + \sum_j p_{kj} \mu_{jk}.$$

By using the Kronecker delta,

$$(19) \quad \delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{cases}$$

these may be combined as one equation valid for all i :

$$(20) \quad \mu_{ik} + \delta_{ik}\mu_k = 1 + \sum_j p_{ij}\mu_{jk}.$$

Now if π is a stationary distribution, we multiply (20) by π_i and sum over all i to give

$$\sum_i \pi_i \mu_{ik} + \sum_i \pi_i \delta_{ik} \mu_k = 1 + \sum_i \sum_j \pi_i p_{ij} \mu_{jk} = 1 + \sum_j \pi_j \mu_{jk}$$

on using the fact that $\pi = \pi P$. Hence, using (19) in the second sum, we have $\pi_k \mu_k = 1$. Because μ_k is uniquely determined and finite, the required results follow. ■

- (21) **Example: Cube** Suppose that a particle performs a random walk on the vertices of a cube in such a way that when it is at a vertex it is equally likely to move along any one of the three edges that meet there, to a neighbouring vertex. Find the mean recurrence time of each vertex.

Solution The state space can be chosen as $S = \{i: 1 \leq i \leq 8\}$ and the transition probabilities are

$$p_{ij} = \begin{cases} \frac{1}{3} & \text{if } i \text{ and } j \text{ are joined by an edge} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\sum_{i \in S} p_{ij} = 1$, and so the stationary distribution is $\pi_i = \frac{1}{8}$; $1 \leq i \leq 8$. By Theorem 15, $\mu_i = 8$; $1 \leq i \leq 8$. ●

More generally, we note that for any finite regular doubly stochastic Markov chain, all states have the same mean recurrence time. This follows easily from the observation that in the doubly stochastic case, we have

$$(22) \quad \frac{1}{d} = \sum_{i \in S} \frac{1}{d} p_{ij} = \frac{1}{d}.$$

Hence, $\pi_i = d^{-1}$ is a stationary distribution and $\mu_i = d$.

- (23) **Example: Library Books** My local lending library permits me to borrow one book at a time. Each Saturday I go to the library. If I have not finished reading the book I renew it; otherwise, I borrow another. It takes me W_r weeks to read the r th book, where $(W_r; r \geq 1)$ is a sequence of independent random variables that are identically distributed. Let X_n be the number of times that I have renewed the book that I take out of the library on the n th Saturday. Show that X_n is a Markov chain and find its transition matrix P .

Find the stationary distribution of P when W_r is uniformly distributed on $\{1, \dots, d\}$.

Solution Let W_r have mass function $f(k)$ and distribution function $F(k)$. Let R denote the record of borrowings and renewals up to, but not including, the book I am currently reading, and suppose that $X_n = i$. Either I renew it again, so $X_{n+1} = i + 1$, or I borrow

a new one, in which case $X_{n+1} = 0$. Because the W_r are independent and identically distributed,

$$\begin{aligned}
 (24) \quad \mathbf{P}(X_{n+1} = i + 1 | X_n = i, R) &= \mathbf{P}(W_1 \geq i + 1 | W_1 = i) \\
 &= \frac{\mathbf{P}(W_1 \geq i + 1)}{\mathbf{P}(W_1 \geq i)} \quad \text{by conditional probability} \\
 &= \frac{1 - F(i + 1)}{1 - F(i)} = \mathbf{P}(X_{n+1} = i + 1 | X_n = i) \\
 &= p_{i,i+1}; \quad i \geq 0.
 \end{aligned}$$

Otherwise,

$$(25) \quad \mathbf{P}(X_{n+1} = 0 | X_n = i, R) = 1 - p_{i,i+1} = \frac{f(i)}{1 - F(i)}.$$

Hence, X is a Markov chain with transition probabilities given by (24) and (25).

If W_r is uniform on $\{1, \dots, d\}$, then

$$p_{i,i+1} = \frac{d - i - 1}{d - i}; \quad 0 \leq i < d - 1$$

and $p_{d-1,0} = 1$. Hence, any stationary distribution π satisfies

$$(26) \quad \pi_{i+1} = \pi_i \frac{d - i - 1}{d - i}; \quad 0 \leq i < d - 1$$

and $\pi_0 = \pi_{d-1}$. Iterating (26) gives $\pi_{i+1} = \frac{d-i-1}{d} \pi_0$, and because $\sum_i^d \pi_i = 1$, it follows that

$$\pi_{i+1} = \frac{2(d - i - 1)}{d(d + 1)}.$$

9.5 The Long Run

It is natural to speculate about the behaviour of the Markov chain X in the long run, that is, as $n \rightarrow \infty$. As usual, we obtain insight from examples before turning to the general case.

Example 9.2.9 Revisited: Survival Recall that

$$p_{ik}(n) = \begin{cases} p^n & \text{if } k = i + n \\ qp^k & \text{if } 0 \leq k \leq n - 1. \end{cases}$$

Now allowing $n \rightarrow \infty$ shows that

$$p_{ik}(n) \rightarrow qp^k.$$

Notice that $\sum_k qp^k = 1$ and that the collection $\pi_k = qp^k$; $k \geq 0$ is a stationary distribution. To see this, just check that $\pi = \pi \mathbf{P}$ because

$$\pi_0 = p \sum_0^{\infty} \pi_k = q$$

and $\pi_k = p\pi_{k-1}$. This stationary distribution does not depend on the starting point of the chain. ●

Example (9.2.7) Revisited: Simple Random Walk In this case, for all i and k ,

$$\lim_{n \rightarrow \infty} p_{ik}(n) = 0. \quad \bullet$$

Example: Gambler's Ruin This is a simple random walk that stops when it reaches 0 or K , and is therefore a Markov chain. The probability of ruin starting from i is p_i , and we have shown above that the probability of winning from i is $1 - p_i$. Hence, for $0 \leq i \leq K$, as $n \rightarrow \infty$,

$$\begin{aligned} p_{i0}(n) &\rightarrow p_i \\ p_{iK}(n) &\rightarrow 1 - p_i \\ p_{ij}(n) &\rightarrow 0 \quad \text{for } 0 < j < K. \end{aligned}$$

The pair $\{p_i, 1 - p_i\}$ is a stationary distribution, but it depends on the initial state of the chain. ●

These examples illustrate the possibilities and agree with our intuition. Roughly speaking, if a chain can get to absorbing states, then it may eventually be absorbed in one of them; on the other hand, if it has no stationary distribution then the chance of finding it in any given state vanishes in the long run.

The most interesting case, as you might expect, arises when the chain has a unique stationary distribution, and the principal result for a finite chain is then the following theorem.

(1) **Theorem** Let X be regular with transition probabilities p_{ik} . Then, as $n \rightarrow \infty$, for any i and k ,

$$p_{ik}(n) \rightarrow \pi_k > 0.$$

Furthermore,

(i) π_k does not depend on i

(ii) $\sum_{k \in S} \pi_k = 1$

and

(iii) $\pi_k = \sum_{i \in S} \pi_i p_{ik}$,

and so π is the stationary distribution of X . In addition, $\pi_k \mu_k = 1$.

Our proof of Theorem 1 will rely on the idea of *coupling*. This technique has many forms and applications; we use a simple version here. Suppose we run two independent chains X and Y with the same transition matrix, and they first take the same value s at time T , say. Now, as we have shown above, the Markov property is preserved at such first passage times, so given $X_T = Y_T = s$, the further progress of X and Y is independent of their

activities before T . Hence, on the event $T \leq n$ we have

$$(2) \quad \mathbf{P}(X_n = k; T \leq n) = \mathbf{P}(Y_n = k; T \leq n),$$

because given $T = t$ both sides are equal to $p_{sk}(n - t)$. The chains are coupled at T .

Now we can tackle the theorem.

Proof of Theorem 1 Let X and Y be independent regular Markov chains with the same state space S and transition matrix p_{ij} . Let X_0 have mass function

$$\mathbf{P}(X_0 = i) = 1,$$

and Y_0 have the stationary distribution of p_{ij} so that

$$\mathbf{P}(Y_0 = i) = \pi_i.$$

Define the Markov chain

$$W = (X, Y),$$

and let T be the first passage time of W to the set $D = \{(x, y): x = y\}$, namely, $T = \min\{n: X_n = Y_n\}$. Now, by (9.2.11), because X and Y are regular, so is W . Hence, T is finite with probability 1 (and has finite mean) by Theorem 9.3.15.

Now, bearing in mind our preparatory remarks above, we can say

$$(3) \quad \begin{aligned} |p_{ik}(n) - \pi_k| &= |\mathbf{P}(X_n = k) - \mathbf{P}(Y_n = k)| \\ &= |\mathbf{P}(X_n = k, n \geq T) - \mathbf{P}(Y_n = k, n \geq T) \\ &\quad + \mathbf{P}(X_n = k, n < T) - \mathbf{P}(Y_n = k, n < T)| \\ &= |\mathbf{P}(X_n = k, n < T) - \mathbf{P}(Y_n = k, n < T)| \text{ by (2)} \\ &\leq \mathbf{P}(T > n), \end{aligned}$$

where the last inequality follows because $|\mathbf{P}(A \cap B) - \mathbf{P}(A \cap C)| \leq \mathbf{P}(A)$ for any events A , B , and C . Because $\mathbf{P}(T > n) \rightarrow 0$ as $n \rightarrow \infty$, we have

$$p_{ik}(n) \rightarrow \pi_k,$$

as required. The rest of the assertions follow because π is the stationary distribution of X . ■

This is a rather useful result; to find the long term behaviour of the chain, we just solve $\pi = \pi \mathbf{P}$, which gives the limiting distribution of X . Indeed, we know from the results of Section 9.3 that this distribution is approached rather quickly, because from Theorem 9.3.15

$$(4) \quad |p_{ij}(n) - \pi_j| < \mathbf{P}(T > n) < c\lambda^n$$

for some constants $c < \infty$ and $\lambda < 1$.

The probabilities $p_{ij}(n)$ are said to approach π_j *geometrically fast*.

(5) **Example** Let X have state space $\{1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}.$$

From the results of Example 9.2.20, we see that when $0 < \alpha + \beta < 2$, as $n \rightarrow \infty$

$$\begin{aligned} p_{11}(n) &\rightarrow \beta, p_{21}(n) \rightarrow \beta \\ p_{12}(n) &\rightarrow \alpha, p_{22}(n) \rightarrow \alpha. \end{aligned}$$

And, of course,

$$(\beta, \alpha) \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix} = (\beta, \alpha),$$

so $\left(\frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta}\right)$ is the stationary distribution as it must be.

When $\alpha + \beta = 0$, the chain is not irreducible, and when $\alpha + \beta = 2$, the chain is not regular (being periodic). ●

- (6) **Example: Entropy of a Markov Source** Let $X = (X_n; n \geq 1)$ be a collection of jointly distributed random variables, and write $X_n = (X_1, \dots, X_n)$. Recall that in Example 9.2.28, we defined the conditional entropy function

$$H(X_{n+1}|X_n) = -\mathbf{E}[\mathbf{E}(\log f(X_{n+1}|X_n)|X_n)].$$

If $H_X = \lim_{n \rightarrow \infty} H(X_{n+1}|X_n)$ exists, then H_X is said to be the *entropy* or *uncertainty* of X .

Now let X be a regular Markov chain with transition matrix \mathbf{P} . Show that H_X does indeed exist and is given by

$$(7) \quad H_X = - \sum_i \pi_i \sum_k p_{ik} \log p_{ik},$$

where π is the stationary distribution of \mathbf{P} .

Solution In Example 9.2.28, it was shown that for a Markov chain X ,

$$\begin{aligned} (8) \quad H(X_{n+1}|X_n) &= H(X_{n+1}|X_n) \\ &= - \sum_i \sum_k \mathbf{P}(X_{n+1} = k | X_n = i) \times \log(\mathbf{P}(X_{n+1} = k | X_n = i)) \\ &\quad \times \mathbf{P}(X_n = i) \\ &= - \sum_i \sum_k \alpha_i(n) p_{ik} \log p_{ik}. \end{aligned}$$

Now, by Theorem 1, as $n \rightarrow \infty$, $\alpha_i(n) \rightarrow \pi_i$, and therefore taking the limit as $n \rightarrow \infty$ of the right-hand side of (8) gives (7), as required. ●

The basic limit theorem (1) tells us that in the long run the probability of finding the regular chain X in state k converges to π_k , for each $k \in S$. It seems plausible that, also in the long run, the proportion of time that X spends visiting k should converge to π_k . The following theorem shows that a more precise version of this vague statement is indeed true. It may be thought of as a type of weak law of large numbers for Markov chains.

(9) **Theorem** Let X be regular with transition matrix \mathbf{P} and stationary distribution π . Let $V_k(n)$ be the number of visits to the state k by X up to time n . Then for any $\varepsilon > 0$,

$$(10) \quad \mathbf{P} \left(\left| \frac{1}{n+1} V_k(n) - \pi_k \right| > \varepsilon \right) \rightarrow 0, \quad \text{as } n \rightarrow \infty.$$

Proof Some groundwork is required before setting about the proof of (10). Let $I_k(n)$ be the indicator of a visit to k at time n , so

$$I_k(n) = \begin{cases} 1 & \text{if } X_n = k \\ 0 & \text{otherwise.} \end{cases}$$

By the basic property of indicators,

$$(11) \quad \mathbf{E}(I_k(n)) = \mathbf{P}(X_n = k) = \alpha_k(n)$$

and for $m \neq r$,

$$(12) \quad \mathbf{E}(I_k(m)I_k(r)) = \alpha_k(s)p_{kk}(t),$$

where $s = \min\{m, r\}$ and $t = |m - r|$. These indicators will be useful because

$$(13) \quad V_k(n) = \sum_{r=0}^n I_k(r).$$

Now we recall that for some constants c and λ with $1 \leq c < \infty$ and $0 < \lambda < 1$,

$$(14) \quad |\alpha_k(n) - \pi_k| < c\lambda^n$$

and

$$(15) \quad |p_{ik}(n) - \pi_k| < c\lambda^n.$$

At last, we are in a position to tackle (10). By Chebyshev's inequality,

$$\begin{aligned} & \mathbf{P} \left(\left| \frac{1}{n+1} V_k(n) - \pi_k \right| > \varepsilon \right) \\ & \leq \mathbf{E} \left(\left(\frac{V_k(n) - (n+1)\pi_k}{(n+1)\varepsilon} \right)^2 \right) \\ & = \frac{1}{(n+1)^2 \varepsilon^2} \mathbf{E} \left(\left(\sum_{r=0}^n (I_k(r) - \pi_k) \right)^2 \right) \\ & = \frac{1}{(n+1)^2 \varepsilon^2} \mathbf{E} \left(\sum_{m=0}^n \sum_{r=0}^n (I_k(m)I_k(r) - \pi_k I_k(m) - \pi_k I_k(r) + \pi_k^2) \right) \\ & = \frac{1}{(n+1)^2 \varepsilon^2} \sum_{m,r} ((\alpha_k(s) - \pi_k)(p_{kk}(t) - \pi_k) \\ & \quad + \pi_k[(\alpha_k(s) - \pi_k) + (p_{kk}(t) - \pi_k) - (\alpha_k(m) - \pi_k) - (\alpha_k(r) - \pi_k)]) \\ & \hspace{15em} \text{by (11) and (12)} \\ & \leq \frac{1}{(n+1)^2 \varepsilon^2} \sum_{m,r} 2c^2(\lambda^s + \lambda^t) \quad \text{by (14) and (15)} \\ & \rightarrow 0 \end{aligned}$$

as $n \rightarrow \infty$, establishing (10). ■

(16) **Corollary** For any bounded function $g(x)$, and any $\epsilon > 0$

$$(17) \quad \mathbf{P} \left(\left| \frac{1}{n+1} \sum_{r=0}^n g(X_r) - \sum_{k \in S} \pi_k g(k) \right| > \epsilon \right) \rightarrow 0,$$

as $n \rightarrow \infty$.

Proof The key to this lies in the observation that

$$\sum_{r=0}^n g(X_r) = \sum_{k \in S} g(k) V_k(n).$$

Hence, we can rewrite (17) as

$$\mathbf{P} \left(\left| \sum_{\substack{k \in S \\ g(k) \neq 0}} g(k) \left(\frac{V_k(n)}{n+1} - \pi_k \right) \right| > \epsilon \right) \leq \sum_{\substack{k \in S \\ g(k) \neq 0}} \mathbf{P} \left(\left| \frac{V_k(n)}{n+1} - \pi_k \right| > \frac{\epsilon}{dg(k)} \right) \rightarrow 0,$$

as $n \rightarrow \infty$, by Theorem 9 (using the fact that S is finite). ■

We can give an immediate application of these results.

(18) **Example: Asymptotic Equipartition for a Markov Source** Let the regular Markov chain X with transition matrix \mathbf{P} and stationary distribution π represent the output from a Markov information source, as defined in Example 9.1.5. Let $X_n = (X_0, \dots, X_n)$ have joint mass function $f(x_0, \dots, x_n) = \mathbf{P}(X_0 = x_0, \dots, X_n = x_n)$, and recall from Example 9.5.6 that the entropy of X is

$$(19) \quad H_X = - \sum_{i \in S} \sum_{k \in S} \pi_i p_{ik} \log p_{ik}.$$

Show that, for any $\delta > 0$, as $n \rightarrow \infty$,

$$(20) \quad \mathbf{P} \left(\left| H_X + \frac{1}{n} \log f(X_0, \dots, X_n) \right| > \delta \right) \rightarrow 0.$$

Solution First, from Example 9.4.8, the sequence $Y_n = \{X_n, X_{n+1}\}; n \geq 0$, is a Markov chain with stationary distribution $(\pi_i p_{ik}; i \in S, k \in S)$. Second, we have

$$(21) \quad \begin{aligned} -\frac{1}{n} \log f(X_0, \dots, X_n) &= -\frac{1}{n} \log(p_{X_0 X_1} p_{X_1 X_2}, \dots, p_{X_{n-1} X_n}) \\ &= -\frac{1}{n} \sum_{r=0}^{n-1} \log p_{X_r X_{r+1}}. \end{aligned}$$

Finally, we note that if we set $g(Y_n) = \log p_{X_n X_{n+1}}$, then Corollary 16 applied to the Markov chain Y shows that

$$\mathbf{P} \left(\left| \frac{1}{n} \sum_{r=0}^{n-1} \log p_{X_r X_{r+1}} - \sum_{i,k} \pi_i p_{ik} \log p_{ik} \right| > \delta \right) \rightarrow 0$$

and (20) follows immediately, on remembering (19) and (21). ●

Here is a useful application of the asymptotic equipartition example. If d is the size of the alphabet, then the total number of messages of length n which the source can emit is d^n . Let us divide them into disjoint sets T and A , where

$$(22) \quad T = \left\{ (x_1, \dots, x_n): \left| \frac{1}{n} \log f(x_1, \dots, x_n) + H_X \right| < \delta \right\}$$

and

$$A = \left\{ (x_1, \dots, x_n): \left| \frac{1}{n} \log f(x_1, \dots, x_n) + H_X \right| \geq \delta \right\}.$$

By Example 18, for any $\epsilon > 0$ and $\delta > 0$, there exists $n_0 < \infty$ such that

$$(23) \quad \mathbf{P}\{(X_1, \dots, X_{n_0}) \in T\} \geq 1 - \epsilon;$$

because this is arbitrarily near 1, sequences in T are called *typical*. Also, by Example 18,

$$\mathbf{P}\{(X_1, \dots, X_{n_0}) \in A\} \leq \epsilon,$$

which is arbitrarily small, so sequences in A are called *atypical*. If you are seeking efficient transmission of messages, it therefore makes sense to concentrate on the typical sequences. It follows that a natural question is, how many typical sequences are there? At this point, we recall that by convention the logarithms in Example 18 are taken to base 2. Hence, from (22),

$$2^{-n(H_X + \delta)} < f(x_1, \dots, x_n) < 2^{-n(H_X - \delta)}.$$

But also, from (23),

$$1 - \epsilon \leq \sum_{x_n \in T} f(x_1, \dots, x_n) \leq 1.$$

Hence, the number $|T|$ of sequences in T satisfies

$$(1 - \epsilon)2^{n(H_X - \delta)} \leq |T| \leq 2^{n(H_X + \delta)},$$

which is to say that, roughly speaking, there are about 2^{nH_X} typical messages of length n .

We end this section with a brief look at Markov chains in general.

Up to now, we have dealt chiefly with finite regular chains because such chains have useful and elegant properties with elementary proofs. However, as some examples have indicated, many chains are irregular or infinite or both. We therefore give a brief account of some of the important results for more general chains; the proofs are all omitted.

In the above sections, it was found that in a finite regular chain, any state d has a recurrence time T_d , which is finite with probability 1 and has finite expectation. When X has countably infinite state space, this need no longer be true, as a glance at the unrestricted simple random walk shows immediately. We therefore distinguish these cases.

(24) **Definition** Let the state d have recurrence time T_d . Then:

(i) If $\mathbf{P}(T_d < \infty) < 1$, then d is said to be transient.

(ii) If $\mathbf{P}(T_d < \infty) = 1$ but $\mathbf{E}(T_d) = \infty$, then d is said to be recurrent null (or persistent null). Otherwise, d is recurrent (or persistent). ▲

These new types of behaviour seem to complicate matters, but the following theorem helps to simplify them again.

(25) **Decomposition Theorem** The state space S can be uniquely partitioned as $S = T \cup C_1 \cup C_2 \cup \dots$, where T is the set of transient states and each C_i is an irreducible closed set of recurrent states.

This means that eventually the chain ends up in some one of the C_i and never leaves it, or it remains forever in the transient states.

Of course, if there is only one closed set of recurrent states matters are even simpler, so the following theorem is useful.

(26) **Theorem** The chain X has a unique stationary distribution π if and only if S contains exactly one recurrent nonnull irreducible subchain C . For each i in C ,

$$(27) \quad \pi_i = \mu_i^{-1},$$

where μ_i is the mean recurrence time of i , and for $i \notin C$, $\pi_i = 0$.

There is also a limit theorem for general Markov chains.

(28) **Theorem** For any aperiodic state k of a Markov chain,

$$p_{kk}(n) \rightarrow \frac{1}{\mu_k} \quad \text{as } n \rightarrow \infty,$$

where the limit is zero if k is null or transient. If i is any other state of the chain, then

$$p_{ik}(n) \rightarrow \frac{1}{\mu_k} \mathbf{P}(T_{ik} < \infty).$$

9.6 Markov Chains with Continuous Parameter

We have suggested above that Markov chains can provide a good description of various real systems. However, a moment's thought about real systems is sufficient to see that many of them do not change their state at integer times. Components fail, your telephone rings, meteorites fall, at any time.

Spurred by this, it is natural to want to study collections of random variables of the form

$$X = (X(t); t \geq 0).$$

Here, $t \in \mathbb{R}$ is often regarded as the time, and then $X(t) \in \mathbb{Z}$ is regarded as the state of the system X at time t . Such a collection is often called a random process. The most obvious thing about X is that it is an uncountable collection of random variables, and it follows that a rigorous account of the behaviour of $X(t)$ is beyond the scope of an elementary text such as this. However, we can discover quite a lot informally in special cases.

First, the remarks in Section 9.1 that motivated our interest in Markov chains apply equally well in continuous time. We therefore make the following definition, analogous to Definition 9.1.1. As usual, $X(t) \in S$, where S is a subset of the integers called the state space.

- (1) **Definition** *The process $X = (X(t); t \geq 0)$ taking values in S is a Markov process (or has the Markov property), if*

$$\mathbf{P}(X(t) = k | X(t_1) = i_1, \dots, X(t_n) = i_n) = \mathbf{P}(X(t) = k | X(t_n) = i_n)$$

for all possible k, i_1, \dots, i_n , and any sequence $0 \leq t_1 < t_2 < \dots < t_n < t$ of times. We write

$$\mathbf{P}(X(t+s) = k | X(s) = i) = p_{ik}(t). \quad \blacktriangle$$

As in the discrete case, $(p_{ik}(t); i, k \in S)$ are known as the transition probabilities, and they satisfy the Chapman–Kolmogorov equations, as follows.

- (2) **Theorem** *For any $s > 0$ and $t > 0$ and $i, k \in S$,*

$$p_{ik}(s+t) = \sum_{j \in S} p_{ij}(s)p_{jk}(t).$$

Proof By the same arguments as we used in Theorem 9.2.12.

- (3)
$$\begin{aligned} p_{ik}(s+t) &= \sum_{j \in S} \mathbf{P}(X(s+t) = k, X(s) = j | X(0) = i) \\ &= \sum_{j \in S} \mathbf{P}(X(t+s) = k | X(s) = j, X(0) = i) \mathbf{P}(X(s) = j | X(0) = i) \\ &= \sum_j p_{ij}(s)p_{jk}(t). \end{aligned} \quad \blacktriangle$$

Given this collection of equations, it is possible to set about solving them in special cases, without any further ado. In fact, we do just that in the next section, but as usual there are a few preliminaries.

First we must ask, do any nontrivial Markov processes exist? (Obviously, the trivial process $X(t) = 1$ for all t is a Markov process, but not a very exciting one.) This question is not as stupid as it may appear to you. Recall that we defined Markov chains by visualizing a counter or particle moving around the vertices of a graph according to some specified distributions, and if necessary we could actually do it. Here, we have started with a collection of probabilities, with no description of how we might actually produce a sequence $X(t)$ having these transition probabilities and joint distributions.

Of course, the answer to the above question is, yes they do exist, and you have already met one, namely, the Poisson process. This was defined by construction in Definition 8.8.1, and we showed that it had the Markov property in Exercise 8.17.3. Henceforth, where necessary, we assume without proof that the processes we consider exist; in more advanced texts, it is shown that they do.

- (4) **Example: Poisson Process** It has already been shown that if $N = (N(t); t \geq 0)$ is a Poisson process with parameter λ , then $N(t) - N(0)$ has a Poisson distribution with parameter λt , which is to say that

$$p_{ik}(t) = \frac{e^{-\lambda t} (\lambda t)^{k-i}}{(k-i)!}.$$

Hence, we can calculate

$$\begin{aligned} \sum_{j=i}^k p_{ij}(s) p_{jk}(t) &= \sum_{j=i}^k \frac{e^{-\lambda s} (\lambda s)^{j-i}}{(j-i)!} \frac{e^{-\lambda t} (\lambda t)^{k-j}}{(k-j)!} \\ &= \frac{e^{-\lambda(t+s)}}{(k-i)!} \sum_{r=0}^{k-i} \binom{k-i}{r} (\lambda s)^r (\lambda t)^{k-i-r} \\ &= \frac{e^{-\lambda(t+s)} (\lambda(s+t))^{k-i}}{(k-i)!} = p_{ik}(s+t). \end{aligned}$$

Thus, the transition probabilities of the Poisson process satisfy the Chapman–Kolmogorov equations, as they must of course by Theorem 2. ●

Note that this result of itself does *not* show that N is a Markov process: there are processes that are not Markov, whose transition probabilities nevertheless satisfy (3). The crucial property which makes N a Markov process is the exponential distribution of times between events. This property is in fact characteristic of Markov processes in general; they wait in each successive state for an exponentially distributed time before moving to the next. Naturally, it is the lack-of-memory property of the exponential distribution that is basically responsible for this essential role in the theory of Markov processes. However, we can do no more here than state the fact baldly; exploring its ramifications is beyond our scope.

One example will suffice to give some trivial insight into these remarks.

Example: Falling Off a Log Let $X(t)$ be a Markov chain with two states 0 and 1. Suppose that transitions from 1 to 0 are impossible. Let us consider transitions from 0 to 1. Because $X(t)$ is Markov, the transition probabilities satisfy the Chapman–Kolmogorov equations. Hence, as $p_{10}(t) = 0$, we have

$$p_{00}(s+t) = p_{00}(s)p_{00}(t) + p_{01}(s)p_{10}(t) = p_{00}(s)p_{00}(t).$$

However, as we have remarked previously, the only bounded solutions to the equation $f(x+y) = f(x)f(y)$ are of the form $f(x) = e^{-\lambda x}$. Hence,

$$p_{00}(t) = e^{-\lambda t}$$

for some $\lambda \geq 0$. The exponential density is forced upon us by the assumption that $X(t)$ is Markov.

9.7 Forward Equations: Poisson and Birth Processes

It is all very well to verify that a previously obtained solution satisfies (9.6.3). A pressing question is, can we solve (9.6.3) without already knowing the answer? We therefore develop a technique for tackling the Chapman–Kolmogorov equations in this section.

First, we observe that for the Poisson process, as $t \rightarrow 0$

$$(1) \quad p_{k,k+1}(t) = \mathbf{P}(N(t) = 1) = \lambda t e^{-\lambda t} = \lambda t + o(t);^\dagger$$

$$(2) \quad p_{kk}(t) = \mathbf{P}(N(t) = 0) = e^{-\lambda t} = 1 - \lambda t + o(t);$$

for $j < k$,

$$(3) \quad p_{kj}(t) = \mathbf{P}(N(t) < 0) = 0;$$

and for $j > k + 1$,

$$(4) \quad p_{kj}(t) = \mathbf{P}(N(t) > 1) = o(t).$$

Equations (1)–(4) say that:

$$(5) \quad N(t) \text{ is nondecreasing.}$$

$$(6) \quad \text{The probability of an event in } [s, s + t] \text{ is proportional to } t, \text{ for small } t, \text{ and does not depend on previous events.}$$

$$(7) \quad \text{The probability of two or more events in } [s, s + t], \text{ for small } t, \text{ is } o(t).$$

What we are going to do now is to seek a Markov process $X(t)$ with transition probabilities $p_{ik}(t)$, which satisfy (5), (6), and (7).

Because $p_{ik}(t)$ satisfies (9.7.3), we have for small t

$$\begin{aligned} p_{ik}(s+t) &= \sum_{j=1}^k p_{ij}(s)p_{jk}(t) \\ &= p_{ik}(s)(1 - \lambda t + o(t)) + p_{i,k-1}(s)(\lambda t + o(t)) + \sum_{j=i}^{k-2} p_{ij}(s)o(t). \end{aligned}$$

Hence,

$$\frac{p_{ik}(s+t) - p_{ik}(s)}{t} = -\lambda p_{ik}(s) + \lambda p_{i,k-1}(s) + o(1)$$

and allowing $t \rightarrow 0$ gives

$$(8) \quad \frac{d}{ds} p_{ik}(s) = -\lambda p_{ik}(s) + \lambda p_{i,k-1}(s),$$

valid for all $0 \leq i \leq k$ [remembering that $p_{i,i-1}(s) = 0$].

[†] We discussed the $o(\cdot)$ notation in Section 7.5.

At $t = 0$, we have the initial condition

$$(9) \quad p_{ii}(0) = 1.$$

The equations (8), as i and k range over all possible values, are known as the *forward equations* for $p_{ik}(t)$, and may be solved in various ways.

Theorem *The solution of (8) is given by*

$$(10) \quad p_{ik}(t) = \frac{e^{-\lambda t} (\lambda t)^{k-i}}{(k-i)!},$$

namely, the transition probabilities of the Poisson process.

Proof We give two methods of proof. First, solve (8) with $k = i$, using (9), to find $p_{ii}(t) = e^{-\lambda t}$. Substituting this into (8) with $k = i + 1$ yields

$$p_{i,i+1}(t) = \lambda t e^{-\lambda t}.$$

A simple induction now yields (10). ■

A second method relies on the generating function

$$G(z, t) = \sum_{k=i}^{\infty} p_{ik}(t) z^k = \mathbf{E}(z^{N(t)} | N(0) = i).$$

Multiply (8) by z^k and sum over k to obtain

$$(11) \quad \frac{\partial G}{\partial t} = \lambda(z-1)G.$$

From (9), we have

$$(12) \quad G(z, 0) = z^i.$$

The solution of (11) that satisfies (12) is

$$(13) \quad G(z, t) = z^i \exp(\lambda t(z-1))$$

and the coefficient of z^k in this expression is just (10). ■

The point of this elaborate reworking is that the simple assumptions (5), (6), and (7) also lead to the Poisson process in a simple and straightforward way. It turns out that a great many useful processes can be analysed by specifying $p_{ik}(t)$ for small t , and all i and k , then obtaining the forward equations, and finally (occasionally) solving them.

(14) Example: The Simple Birth Process A population of individuals grows as follows. Each member of the population in existence at time t may be replaced by two new individuals during $[t, t+h]$ with probability $\lambda h + o(h)$, independently of the other members of the population. Otherwise, the given member of the population remains intact during

$[t, t + h]$ with probability $1 - \lambda h + o(h)$, also independently of the rest of the population. If the population at time t is $X(t)$, and $X(0) = 1$, show that

$$(15) \quad \mathbf{E}(z^{X(t)}) = \frac{z}{z + (1 - z)e^{\lambda t}}.$$

What is $\mathbf{E}(X(t))$?

Solution Let $p_n(t) = \mathbf{P}(X(t) = n)$ and

$$p_{jk}(t) = \mathbf{P}(X(t) = k | X(0) = j).$$

Suppose that $X(t) = i$. Because each individual is replaced (or not) independently of all the others we have, as $h \rightarrow 0$,

$$\begin{aligned} p_{ii}(h) &= (1 - \lambda h + o(h))^i \\ p_{i,i+1}(h) &= i(\lambda h + o(h))(1 - \lambda h + o(h))^{i-1} = i\lambda h + o(h) \\ p_{ik}(h) &= o(h); \quad k > i + 1 \\ p_{ik}(h) &= 0; \quad k < i. \end{aligned}$$

Following the by now familiar routine, we find

$$p_k(t + h) = (1 - \lambda kh)p_k(t) + \lambda(k - 1)hp_{k-1}(t) + o(h)$$

and so

$$(16) \quad \frac{\partial}{\partial t} p_k(t) = -\lambda k p_k(t) + \lambda(k - 1)p_{k-1}(t).$$

Now we set $G_X(z, t) = \mathbf{E}(z^{X(t)})$, and notice that because probability generating functions are differentiable, at least for $|z| < 1$, we have

$$(17) \quad \frac{\partial G_X}{\partial z} = \sum_{k=1}^{\infty} k p_k(t) z^{k-1}.$$

Now, on multiplying (16) by z^k and summing over k , we notice $\frac{\partial G_X}{\partial z}$ appearing on the right-hand side. In fact,

$$(18) \quad \frac{\partial G_X}{\partial t} = \lambda z(z - 1) \frac{\partial G_X}{\partial z}.$$

Also, because $X(0) = 1$,

$$(19) \quad G_X(z, 0) = z.$$

By inspection, for any differentiable function $h(\cdot)$, the function

$$(20) \quad G(z, t) = h\left(\lambda t + \int^z \frac{1}{v(v-1)} dv\right)$$

satisfies (18). Imposing the boundary condition (19) reveals that

$$z = h \left(\int^z \frac{1}{v(v-1)} dv \right) = h \left(\log \left(\frac{z-1}{z} \right) \right).$$

Hence, the function $h(\cdot)$ is given by $h(y) = (1 - e^y)^{-1}$ and so

$$G_X(z, t) = \frac{1}{1 - \exp \left(\lambda t + \log \left(\frac{z-1}{z} \right) \right)} = \frac{z}{z + (1-z)e^{\lambda t}},$$

as required.

Now, to find $\mathbf{E}(X(t))$, we have a choice of methods. Obviously, by differentiating (15) with respect to z and setting $z = 1$, we find

$$(21) \quad \mathbf{E}(X(t)) = \frac{\partial}{\partial z} G_X(1, t) = e^{\lambda t}.$$

However, we could have obtained $\mathbf{E}(X(t))$ without solving (18). If we assume $\mathbf{E}[(X(t))^2]$ exists, then differentiating (18) with respect to z and setting $z = 1$ yields

$$\frac{\partial}{\partial t} \frac{\partial G_X}{\partial z}(1, t) = \lambda \frac{\partial G_X}{\partial z}(1, t).$$

This has solution given by (21). Or we could simply have noted that X has a geometric distribution with parameter $e^{-\lambda t}$. ●

9.8 Forward Equations: Equilibrium

Guided by our glances at the Poisson and simple birth processes, we can now outline a simple technique for dealing with some elementary Markov chains. The aim is to obtain forward equations, so from the Chapman–Kolmogorov equations we write, for $h > 0$,

$$(1) \quad \frac{p_{ik}(t+h) - p_{ik}(t)}{h} = \frac{1}{h} \left(\sum_j p_{ij}(t) p_{jk}(h) - p_{ik}(t) \right).$$

We want to let $h \downarrow 0$. By inspection of (1), this is possible if for some finite numbers $(g_{jk}; j, k \in S)$ we have as $h \rightarrow 0$,

$$(2) \quad p_{jk}(h) = g_{jk}h + o(h),$$

and

$$(3) \quad p_{kk}(h) = 1 + g_{kk}h + o(h).$$

In this case, we obtain the required forward equations by letting $h \downarrow 0$ in (1), to give

$$(4) \quad \frac{\partial}{\partial t} p_{ik}(t) = \sum_j p_{ij}(t) g_{jk}.$$

The application of this idea is best illustrated by examples. We give perhaps the simplest here, others follow in due course.

- (5) **Example: Machine** A machine can be either up or down. [You can interpret this figuratively (working/not working), or literally (a lift), it makes no difference to the mathematics.] If it is up at time t , then it goes down during $[t, t + h]$ with probability $\alpha h + o(h)$, independently of its past record. Otherwise, it stays up with probability $1 - \alpha h + o(h)$ during $[t, t + h]$. Likewise, if it is down at time t , it goes up in $[t, t + h]$ with probability $\beta h + o(h)$, independently of its past, or it stays down with probability $1 - \beta h + o(h)$.
- (a) If it is up at $t = 0$, find the probability that it is down at time $t > 0$.
- (b) Let $N(t)$ be the number of occasions on which it has gone down during $[0, t]$. Find $\mathbf{E}(N(t))$.
- (c) Find the probability generating function $\mathbf{E}(z^{N(t)})$.

Solution (a) Let $X(t)$ be the state of the machine, where $X(t) = 0$ if it is up at t and $X(t) = 1$ if it is down at t . By the assumptions of the question, $X(t)$ is a Markov process, and

$$\begin{aligned} p_{01}(t+h) &= p_{00}(t)\alpha h + p_{01}(t)(1 - \beta h) + o(h) \\ p_{00}(t+h) &= p_{01}(t)\beta h + p_{00}(t)(1 - \alpha h) + o(h). \end{aligned}$$

Hence,

$$(6) \quad \frac{d}{dt} p_{01}(t) = -\beta p_{01}(t) + \alpha p_{00}(t)$$

$$(7) \quad \frac{d}{dt} p_{00}(t) = -\alpha p_{00}(t) + \beta p_{01}(t)$$

and, because $X(0) = 0$,

$$(8) \quad p_{00}(0) = 1.$$

Solving (6), (7), and (8) gives the required probability of being down at t

$$p_{01}(t) = \frac{\alpha}{\alpha + \beta} (1 - e^{-(\alpha + \beta)t}).$$

(b) The first thing to realize here is that $N(t)$ is *not* a Markov process. However, if we let $M(t)$ be the number of times the machine has gone up during $[0, t]$, then

$$\mathbf{Y}(t) = \{N(t), M(t)\}$$

is a Markov process. By the assumptions of the problem, as $h \rightarrow 0$,

$$\begin{aligned} \mathbf{P}(\mathbf{Y}(t+h) = (k+1, k) | \mathbf{Y}(t) = (k, k)) &= \alpha h + o(h) \\ \mathbf{P}(\mathbf{Y}(t+h) = (k, k) | \mathbf{Y}(t) = (k, k)) &= 1 - \alpha h + o(h) \\ \mathbf{P}(\mathbf{Y}(t+h) = (k, k) | \mathbf{Y}(t) = (k, k-1)) &= \beta h + o(h) \\ \mathbf{P}(\mathbf{Y}(t+h) = (k, k-1) | \mathbf{Y}(t) = (k, k-1)) &= 1 - \beta h + o(h). \end{aligned}$$

Hence, if $f_{kj}(t) = \mathbf{P}(Y(t) = (k, j))$, the forward equations may be derived routinely as

$$(9) \quad \frac{d}{dt} f_{kk}(t) = -\alpha f_{kk}(t) + \beta f_{k,k-1}(t); \quad k \geq 0.$$

$$(10) \quad \frac{d}{dt} f_{k,k-1}(t) = -\beta f_{k,k-1}(t) + \alpha f_{k-1,k-1}(t); \quad k \geq 1,$$

where $f_{0,-1}(t) = 0$.

Now consider a Poisson process $Z(t)$ of rate α . By construction of $N(t)$, $\mathbf{P}(N(t) = k) \leq \mathbf{P}(Z(t) = k)$ for all k . Hence, $\mathbf{E}(N(t))$ exists because it is less than $\mathbf{E}(Z(t)) = \alpha t$. In fact,

$$\mathbf{P}(N(t) = k) = f_{kk}(t) + f_{k,k-1}(t),$$

and so $\mathbf{E}(N(t)) = m_1(t) + m_2(t)$, where

$$m_1(t) = \sum_{k=1}^{\infty} k f_{kk}(t)$$

and

$$m_2(t) = \sum_{k=1}^{\infty} k f_{k,k-1}(t).$$

Multiplying (9) and (10) by k , and summing gives

$$(11) \quad \frac{dm_1}{dt} = -\alpha m_1 + \beta m_2$$

and

$$(12) \quad \begin{aligned} \frac{dm_2}{dt} &= -\beta m_2 + \alpha \sum_k (k-1+1) f_{k-1,k-1}(t) \\ &= -\beta m_2 + \alpha m_1 + \alpha \sum_k f_{k-1,k-1}(t) \\ &= -\beta m_2 + \alpha m_1 + \alpha \mathbf{P}(N(t) = M(t)) \\ &= -\beta m_2 + \alpha m_1 + \alpha p_{00}(t) \end{aligned}$$

because the machine is up if $N(t) = M(t)$. Hence, adding (11) and (12), we have

$$\frac{d}{dt} \mathbf{E}(N(t)) = \alpha p_{00}(t),$$

and so, using the result of (a),

$$\mathbf{E}(N(t)) = \int_0^t \alpha \left(\frac{\beta}{\alpha + \beta} + \frac{\alpha e^{-(\alpha+\beta)v}}{\alpha + \beta} \right) dv.$$

It follows that as $t \rightarrow \infty$, $t^{-1} \mathbf{E}(N(t)) \rightarrow \alpha\beta/(\alpha + \beta)$.

(c) Let

$$x(t, z) = \sum_{k=0}^{\infty} f_{kk}(t) z^k \quad \text{and} \quad y(t, z) = \sum_{k=1}^{\infty} f_{k,k-1}(t) z^k.$$

Then $\mathbf{E}(z^{N(t)}) = x(t, z) + y(t, z)$, and multiplying each of (9) and (10) by z^k and summing over k gives

$$\frac{\partial x}{\partial t} = -\alpha x + \beta zy, \quad \text{and} \quad \frac{\partial y}{\partial t} = -\beta y + \alpha zx.$$

This pair of simultaneous differential equations is solved by elementary methods, subject to the initial conditions $x(0, z) = 1$ and $y(0, z) = 0$, to yield

$$x(t, z) = [(\alpha + \lambda_2(z))/(\lambda_2(z) - \lambda_1(z))]e^{\lambda_1(z)t} + [(\alpha + \lambda_1(z))/(\lambda_1(z) - \lambda_2(z))]e^{\lambda_2(z)t}$$

and

$$y(t, z) = (\alpha + \lambda_1)(\alpha + \lambda_2)(e^{\lambda_1(z)t} - e^{\lambda_2(z)t})/(\beta z(\lambda_2 - \lambda_1)),$$

where

$$\lambda_1(z) = \frac{1}{2}[-(\alpha + \beta) + ((\alpha - \beta)^2 + 4\alpha\beta z^2)^{\frac{1}{2}}]$$

and

$$\lambda_2(z) = \frac{1}{2}[-(\alpha + \beta) - ((\alpha - \beta)^2 + 4\alpha\beta z^2)^{\frac{1}{2}}]. \quad \bullet$$

Recalling our results about chains in discrete time, it is natural to wonder whether chains in continuous time have stationary distributions and whether $p_{ij}(t)$ converges as $t \rightarrow \infty$. A detailed answer to these questions is far beyond our scope, but we can make some guarded remarks. Let us start with the simplest.

- (13) **Theorem** *Let $X(t)$ be a finite Markov process with transition matrix $p_{ij}(t)$. Then $\lim_{t \rightarrow \infty} p_{ij}(t)$ exists for all i and j .
If $X(t)$ is irreducible then the limit is independent of i , we write*

$$\lim_{t \rightarrow \infty} p_{ij}(t) = \pi_j.$$

Furthermore, π_j satisfies $\sum_j \pi_j = 1$ and

$$\pi_j = \sum_i \pi_i p_{ij}(t); \quad t \geq 0,$$

π is the stationary distribution of $X(t)$.

[A chain is irreducible if for each i, j there is some finite t such that $p_{ij}(t) > 0$.]

- (14) **Example 5 Revisited** In this case, we have

$$p_{01}(t) = \frac{\alpha}{\alpha + \beta}(1 - e^{-(\alpha + \beta)t}) \rightarrow \begin{cases} \frac{\alpha}{\alpha + \beta} & \text{if } \alpha + \beta > 0 \\ 0 & \text{otherwise,} \end{cases}$$

with three similar results for p_{00} , p_{10} , and p_{11} . If $\alpha = \beta = 0$, then $X(t) = X(0)$ for all t . However, if $\alpha\beta > 0$, then the chain is irreducible and has stationary distribution $\pi =$

$(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$. We can check that for all t

$$\begin{aligned} \pi_0 &= \frac{\beta}{\alpha + \beta} \\ &= \frac{\beta}{\alpha + \beta} \left(\frac{\beta}{\alpha + \beta} + \frac{\alpha e^{-(\alpha+\beta)t}}{\alpha + \beta} \right) + \frac{\alpha}{\alpha + \beta} \left(\frac{\beta}{\alpha + \beta} - \frac{\beta e^{-(\alpha+\beta)t}}{\alpha + \beta} \right) \\ &= \pi_0 p_{00}(t) + \pi_1 p_{10}(t). \end{aligned}$$

In practice, the state space is often countably infinite, and of course we would like to use the forward equations (4). The following theorem is relevant.

(15) **Theorem** *Let $X(t)$ be an irreducible Markov process with transition matrix $p_{ij}(t)$, satisfying (2) and (3) above. Then $\lim_{t \rightarrow \infty} p_{ij}(t)$ exists and is independent of i , for all j .*

There are two possibilities.

Either (a) $\lim_{t \rightarrow \infty} p_{ij}(t) = \pi_j > 0$,

where $\sum \pi_i = 1$ and

$$\pi_j = \sum_i \pi_i p_{ij}(t)$$

and

(16)
$$\sum_i \pi_i g_{ij} = 0$$

for all j ;

or (b) $\lim_{t \rightarrow \infty} p_{ij}(t) = 0$.

We give no proof of this result, but you may notice with some pleasure that there are no tiresome reservations about periodicity.

(17) **Example: Queue** Let $X(t)$ be the length of queue formed before a single service point at time t . The times between arrivals are exponentially distributed with parameter λ ; each individual is served on reaching the head of the queue; each service time is exponentially distributed with parameter μ ; interarrival times and service times are all independent of each other.

It follows that as $h \rightarrow 0$, when $X(t) > 0$,

$$\begin{aligned} \mathbf{P}(X(t+h) - X(t) = 1) &= \mathbf{P}(\text{one arrival; no service completed in } (t, t+h)) \\ &= \lambda h(1 - \mu h) + o(h) = \lambda h + o(h). \end{aligned}$$

Likewise,

$$\mathbf{P}(X(t+h) - X(t) = -1) = \mu h + o(h)$$

and

$$\mathbf{P}(X(t+h) - X(t) = 0) = 1 - (\lambda + \mu)h + o(h).$$

The process $X(t)$ is Markov, by the properties of the exponential density, and the above statements show that

$$\begin{aligned} p_{i,i+1}(h) &= \lambda h + o(h), \\ p_{i,i-1}(h) &= \mu h + o(h); \quad i \neq 0 \\ p_{ii}(h) &= 1 - (\lambda + \mu)h + o(h); \quad i \neq 0. \end{aligned}$$

When $i = 0$, no service can be completed so we have $p_{01}(h) = \lambda h + o(h)$ and $p_{00}(h) = 1 - \lambda h + o(h)$. These supply us with all the numbers g_{ij} , and so by (16) to find the stationary distribution we seek a solution to the equations $\pi G = 0$, that is,

$$\begin{aligned} -\lambda\pi_0 + \mu\pi_1 &= 0 \\ \lambda\pi_{i-1} - (\lambda + \mu)\pi_i + \mu\pi_{i+1} &= 0; \quad i \geq 1. \end{aligned}$$

Solving recursively shows that

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \pi_0; \quad i \geq 0,$$

and so a stationary distribution exists if $(\frac{\lambda}{\mu}) < 1$, and it is given by

$$\pi_i = \left(1 - \frac{\lambda}{\mu}\right) \left(\frac{\lambda}{\mu}\right)^i. \quad \bullet$$

9.9 The Wiener Process and Diffusions

All the Markov processes that we have considered up to now have been “discrete,” which is to say that they take values in the integers (or some other countable set). They have thus been “jump processes,” in the sense that transitions between discrete states take place instantaneously at times that may be fixed (as in the simple random walk), or they may be random times indexed by a continuous parameter (as in the Poisson process). But it is a matter of simple observation that many real random processes do not make jump transitions in a countable set of states. In complete contrast, such processes are seen to move continuously between their possible states, which typically lie in some interval of the real line (or all the real line). We may mention noisy electronic signals, meteorological data, and the size and position of sunspots, for example.

We have seen that Markov chains have been effective and amenable in describing processes with discrete state space. It is natural to ask if there are also useful processes that are continuous and have the Markov property (Definition 9.6.1). The answer is yes, and by far the simplest and most useful of such models is that called the Wiener process (also known as Brownian motion for reasons we discuss in a moment). Here it is:

- (1) **Definition** The random process $W(t), t \geq 0$, is called the Wiener process if it satisfies
- (a) $W(0) = 0$
 - (b) $W(t)$ is continuous
 - (c) $W(s + t) - W(s)$ is normally distributed with mean 0 and variance $\sigma^2 t$, for all $s, t \geq 0$, and some constant $\sigma^2 > 0$.
- Remark** Note that we usually take $\sigma^2 = 1$, in which case this may be called the standard Wiener process. Our Wiener processes are standard unless otherwise stated.
- (d) $W(t)$ has independent increments, which is to say that for any $0 < t_0 < t_1 < \dots < t_n$,

$$W(t_0), W(t_1) - W(t_0), \dots, W(t_n) - W(t_{n-1}),$$
 are independent random variables.

[Note that it still seems to be an open problem as to whether pairwise independence of increments is sufficient to define $W(t)$.] ▲

This may seem a rather abrupt and arbitrary definition, so our first task must be to justify it. First, of course, we must check that it is indeed a Markov process satisfying Definition (9.6.1). We write, for any $t_0 < t_1 < \dots < t_n < t_n + s$,

$$\begin{aligned} \mathbf{P}(W(t_n + s) \leq w | W(t_n) = w_n, \dots, W(t_0) = w_0) \\ &= \mathbf{P}(W(t_n + s) - w_n \leq w - w_n | W(t_n), W(t_n) - W(t_{n-1}), \dots, W(t_0) = w_0) \\ &= \mathbf{P}(W(t_n + s) - w_n \leq w - w_n | W(t_n) = w_n) \text{ by the independence of increments} \\ &= \mathbf{P}(W(t_n + s) \leq w | W(t_n) = w_n), \end{aligned}$$

which is the Markov property, as required.

The second natural question is, why have we picked that particular set of properties to define $W(t)$? The answer most easily appears from the history and background of the process, which is also of interest in its own right.

Ancient peoples were able to cut and polish rock crystal, and also glass when it was invented. They were aware that, when lens-shaped, such objects distorted observation and focussed light. However, the crucial step of exploiting this commonplace fact for useful purposes was taken in Holland at the end of the sixteenth century. Hans Lippeshey had an effective telescope by 1608, and Hans and Zacharias Janssen developed a microscope at about the same time. Galileo immediately used the telescope to revolutionize astronomy and cosmology, and Robert Hooke and others used the microscope to revolutionize our knowledge of smaller-scale aspects of the universe.

In particular, Antonie van Leeuwenhoek noted tiny objects in drops of water, which he called animalcules. He observed that their motion “was so fast, and so random, upwards, downwards and round in all directions that it was truly wonderful to see.” He attributed this motion to the objects being alive, and indeed he is credited with first identifying bacterial cells.

However, it was the botanist Robert Brown who (in 1822) conducted key experiments that demonstrated that this erratic motion is exhibited by *any* sufficiently small particle, inanimate as well as animate. He began his observations on pollen, which is in general too large to show the effect, but he observed that the smaller particles were in motion. Further, the smaller the particle the more vigorous the motion, and he obtained the same result using “every mineral which I could reduce to powder,” including arsenic and “a fragment of the Sphinx.” The motion must therefore be purely mechanical, and is in fact caused by the ceaseless battering of the atoms of the fluid in which the particles are suspended. What are the properties of the process generated by the movements of such a randomly battered particle?

- (i) First the physical nature of particles and molecules entails that their movements in nonoverlapping time intervals are independent. This is 1(d).
- (ii) Second, the path of a particle is surely continuous. This is 1(b).
- (iii) Third, the position of the particle at time t is the cumulative sum of an arbitrarily large number of small steps that are independent by (i) above. The sizes of steps over time intervals of the same length have the same distribution, by the homogeneous nature of the conditions of the problem. The central limit theorem therefore applies, and we see that the position of the particle $X(t)$ is normally distributed. The mean of $X(t)$ is 0 by symmetry. For the variance, we note that the variance of the position of a discrete random walk is proportional to its duration. Because we envisage $X(t)$ as the continuous limit of such a walk, it is natural to set $\text{var } X(t) = c^2t$. That is, 1(c).
- (iv) Without loss of generality, we start the particle at 0. This is 1(a).

The first successful attempt at a mathematical and scientific description of the effect was undertaken by A. Einstein in 1905. (The same year as his better-known work on relativity.) He characterized the motion in terms of physical laws and constants; a description that later allowed Perrin to obtain Avogadro’s number, and hence a Nobel prize. Another mathematical description is implicit in the earlier work of L. Bachelier in 1900, but this (by contrast) was ignored for half a century.

In both cases, the idea underlying the model is that the effect of the impact of atoms on the particles is to force them to execute a random walk, whose steps (on any scale) are equally likely to be in any direction. Clearly, the steps must be independent, and an application of the central limit theorem tells us that steps in any given direction must be normally distributed. These are just the properties we set out in Definition (1).

There is a third, rather less obvious question about the process $W(t)$ defined in (1). That is, does it exist? This may seem paradoxical, but recall our discussion after (9.6.3). The point is that almost every other process in this book was defined by construction, and then we deduced its joint distributions. Here we simply stated the joint distributions (implicitly); thus leaving open the question of whether a construction is possible. The answer “yes,” together with a construction, was supplied by N. Wiener in a series of papers from 1918. This is why we give it that name. The proof is too intricate for inclusion here, so we move on to look at the basic properties of $W(t)$.

A key point to grasp here is that such properties fall into two types: sample-path properties and distributional properties. The idea of what we mean by that is best conveyed by giving some such properties. Here is a list of properties of the paths of $W(t)$.

With probability 1,

- (a) $W(t)$ is continuous.
- (b) $W(t)$ is not differentiable anywhere.
- (c) $W(t)$ changes sign infinitely often in any interval $[0, \varepsilon]$, $\varepsilon > 0$.
- (d) $W(t)/t \rightarrow 0$, as $t \rightarrow \infty$.

Roughly speaking, $W(t)$ is incredibly spiky, but does not get too far away from zero.

These path properties are not so easy to verify so we turn our attention to the other kind, that is, properties of the joint distributions of $W(t)$, which are again best illustrated by examples.

It is useful to recall that the $N(0, \sigma^2)$ normal density is denoted by

$$\phi_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{x^2}{\sigma^2}\right),$$

and we omit the suffix if $\sigma^2 = 1$, so $\phi_1(x) = \phi(x)$. Also, recall from the end of 8.10, the review of Chapter 8, that the joint distribution of a collection of multivariate normal random variables is determined by their means, variances, and covariances. Hence, we have for the Wiener process

- (2) **Example: Joint Distribution** Because increments are independent and normally distributed, the joint density of $W(t_1), W(t_2), \dots, W(t_n)$, where $t_1 < t_2 < \dots < t_n$, is given by

$$f(w_1, \dots, w_n) = \phi_{t_1}(w_1)\phi_{t_2-t_1}(w_2 - w_1) \dots \phi_{t_n-t_{n-1}}(w_n - w_{n-1}).$$

In particular, the joint density of $W(s)$ and $W(t)$ is

$$(3) \quad f(s, x; t, y) = \frac{1}{2\pi} \frac{1}{\sqrt{s(t-s)}} \exp\left\{-\frac{1}{2} \frac{x^2}{s} - \frac{1}{2} \frac{(y-x)^2}{t-s}\right\}, \quad 0 < s < t.$$

Now by comparison with Example (8.11) we see immediately that

$$\rho(W(s), W(t)) = \sqrt{\frac{s}{t}} \quad \text{and} \quad \text{cov}(W(s), W(t)) = s. \quad \bullet$$

Because $W(t)$ is a Markov process its transition probabilities satisfy the Chapman–Kolmogorov equations, as we showed in Theorems 9.2.12 and 9.6.2 for discrete Markov processes.

In this case, because $W(t)$ has continuous state space, the equations are expressed in terms of an integral. That is to say, by considering the possible values z of $W(u)$ for $s < u < t$, and using the Markov property at u , we have

$$(4) \quad f(s, x; t, y) = \int_{\mathbb{R}} f(s, x; u, z) f(u, z; t, y) dz.$$

In Problem 9.35, you are asked to show that the transition density in (3) does indeed satisfy these equations.

Joint densities can be readily used to find joint moments and conditional densities, but there are often simpler techniques. Here is an example.

- (5) **Example: Conditioning** We could use (2) and (3) to obtain the conditional density of $W(s)$ given $W(t) = y$, as

$$\begin{aligned} f_{W(s)|W(t)}(x|W(t) = y) &= \frac{f(s, x; t, y)}{\phi_t(y)} \\ &\propto \exp \left\{ -\frac{1}{2} \frac{x^2}{s} - \frac{1}{2} \frac{(y-x)^2}{t-s} + \frac{1}{2} \frac{y^2}{t} \right\} \\ &\propto \exp \left\{ -\frac{1}{2} \frac{t}{s(t-s)} \left(x - \frac{sy}{t} \right)^2 \right\}. \end{aligned}$$

By inspection, this is a normal density with mean sy/t and variance $\sqrt{s(t-s)/t}$ (where, for simplicity, we have omitted the normalizing constant).

That is to say, for $0 < s < t$,

$$(6) \quad \mathbf{E}(W(s)|W(t)) = \frac{s}{t} W(t),$$

$$(7) \quad \text{var}(W(s)|W(t)) = s \left(1 - \frac{s}{t} \right).$$

Hence,

$$\begin{aligned} (8) \quad \text{cov}(W(s), W(t)) &= \mathbf{E}(W(s)W(t)) \\ &= \mathbf{E}(\mathbf{E}(W(s)W(t)|W(t))) \\ &= \mathbf{E} \left[\frac{s}{t} W(t)^2 \right] = s = s \wedge t, \end{aligned}$$

as we remarked above.

In fact, in this case and in many other similar situations, it is easier and quicker to recall the clever device from Example (8.20). Thus, we first calculate $\text{cov}(W(s), W(t))$ directly as

$$\begin{aligned} \mathbf{E}(W(s)W(t)) &= \mathbf{E}\{W(s)[W(t) - W(s)] + W(s)^2\} \\ &= \mathbf{E}W(s)^2, \quad \text{by independent increments} \\ &= s = s \wedge t = \min\{s, t\}. \end{aligned}$$

Hence, $W(s)/\sqrt{s}$ and $W(t)/\sqrt{t}$ have a standard bivariate normal density with correlation

$$\rho = \mathbf{E} \left(\frac{W(s)}{\sqrt{s}} \frac{W(t)}{\sqrt{t}} \right) = \sqrt{\frac{s}{t}}.$$

By (8.20), we find immediately that given $W(t)$, $W(s)/\sqrt{s}$ is normal with conditional mean

$$\mathbf{E} \left(\frac{W(s)}{\sqrt{s}} \middle| W(t) \right) = \sqrt{\frac{s}{t}} \frac{W(t)}{\sqrt{t}} = \frac{\sqrt{s}}{t} W(t),$$

and conditional variance

$$\text{var} \left(\frac{W(s)}{\sqrt{s}} \middle| W(t) \right) = 1 - \rho^2 = 1 - \frac{s}{t},$$

as are found above. ●

A famous and important special case is the so-called “Brownian Bridge,” which (perhaps surprisingly) turns out to be useful in making statistical inferences about empirical distribution functions.

- (9) **Example: Brownian Bridge** This is the process $B(t)$, $0 \leq t \leq 1$, defined to be $W(t)$, conditional on the event $W(1) = 0$. (It is sometimes called the “tied-down Wiener process.”) Because $W(t)$ has multivariate normal joint distributions, it also follows immediately that $B(t)$ does. To be precise, from the results of Example (5), we have

$$\mathbf{E}B(t) = \mathbf{E}(W(t)|W(1) = 0) = 0;$$

and for $0 < s < t < 1$,

$$\begin{aligned} \mathbf{E}(B(s)B(t)) &= \mathbf{E}(W(s)W(t)|W(1) = 0) \\ &= \mathbf{E}(\mathbf{E}(W(s)W(t)|W(t), W(1) = 0)|W(1) = 0) \\ &\quad \text{by (5.7.4), the tower property,} \\ &= \mathbf{E}(\mathbf{E}(W(s)W(t)|W(t))|W(1) = 0) \\ &\quad \text{by independence of the increments of } W(t), \\ &= \mathbf{E}\left(\frac{s}{t}W(t)^2|W(1) = 0\right) \quad \text{by (6)} \\ &= \frac{s}{t}t(1-t) \quad \text{by (7)} \\ &= s \wedge t - st. \end{aligned}$$

Obviously, the Brownian Bridge cannot have independent increments because it is forced to satisfy $B(1) = 0$. Nevertheless, it is a Markov process, as you can easily show by verifying that its transition probabilities satisfy (9.6.1). ●

In some ways, it is unsatisfying to have a process defined as a conditional version of another, especially when the conditioning is on an event of probability zero. The following result is therefore very useful.

- (10) **Lemma** Let $B^*(t) = W(t) - tW(1)$, $0 \leq t \leq 1$. Then $B^*(t)$ is also the Brownian Bridge.

Proof $B^*(t)$ has multivariate normal joint distributions because $W(t)$ does; obviously, $B^*(t)$ has zero mean. Also, for $s < t < 1$,

$$\begin{aligned} \text{cov}(B^*(s), B^*(t)) &= \mathbf{E}(W(s)W(t) - sW(t)W(1) - tW(s)W(1) + stW^2(1)) \\ &= s - st - st + st = s \wedge t - st \\ &= \text{cov}(B(s), B(t)). \end{aligned}$$

The proof is complete, when we recall from (8.10) that multinormal distributions are determined by their first and second joint moments. ■

Besides conditional processes, there are several other operations on the Wiener process that have interesting and important results.

- (11) **Example: Scaling** For any constant $c > 0$, the process $W^*(t) = \sqrt{c}W(\frac{t}{c})$ is also a Wiener process.

Solution It is immediate that $W^*(t)$ is continuous, starts at zero, and has independent normal increments because $W(t)$ does.

Finally, we need only check the variance of an increment by calculating

$$\text{var}(W^*(t) - W^*(s)) = c\mathbf{E}\left[W\left(\frac{t}{c}\right) - W\left(\frac{s}{c}\right)\right]^2 = c\left(\frac{t-s}{c}\right) = t - s.$$

Hence, $W^*(t)$ satisfies the conditions to be a Wiener process. ●

- (12) **Example: Ornstein–Uhlenbeck Process** This is obtained from $W(t)$ by another non-linear scaling of time and size; thus,

$$U(t) = e^{-t}W(e^{2t}).$$

This process has been used by physicists as a model for the velocity of a particle in a fluid at time t . The joint distributions of $U(t)$ are multivariate normal because those of $W(t)$ are. In particular, we have

$$\mathbf{E}U(t) = 0$$

and

$$\text{var}U(t) = \mathbf{E}[U(t)]^2 = e^{-2t}\mathbf{E}(W(e^{2t})^2) = 1.$$

Also, for $s, t > 0$,

$$\mathbf{E}(U(t)U(t+s)) = e^{-2t-s}\mathbf{E}[W(e^{2t})W(e^{2(t+s)})] = e^{-s}.$$

Hence, $U(t)$ and $U(t+s)$ have the standard bivariate normal density with correlation e^{-s} ; we can therefore write, if we want,

$$U(s+t) = e^{-s}U(t) + \sqrt{1 - e^{-2s}}Z,$$

where Z is a normal $N(0, 1)$ random variable that is independent of $U(t)$.

You can show that $U(t)$ is a Markov process (exercise). ●

- (13) **Example: Drifting Wiener Process** This is obtained from the standard Wiener process $W(t)$ by setting

$$D(t) = \sigma W(t) + \mu t.$$

If $\sigma W(t)$ represents the position of a particle enjoying Brownian motion in some fluid, then μt may be interpreted as a superimposed global motion of the fluid moving with constant velocity μ . Alternatively, we may regard $D(t)$ as the continuous limit of an asymmetric simple random walk. ●

- (14) **Example: Exponential (or Geometric) Wiener Process** This is obtained from $W(t)$ by setting

$$Y(t) = \exp(\mu t + \sigma W(t)).$$

Because this cannot be negative and $\log Y(t)$ has independent increments, it is popular as a model for stock prices. It is also called geometric Brownian motion. ●

(15) **Example: Reflected Wiener Process** This is obtained from $W(t)$ by setting

$$Z(t) = |W(t)|.$$

It is interpreted as the Wiener process with a “reflecting barrier” at the origin. It can be shown to be a Markov process, with density $2\phi_t(z)$, giving

$$\mathbf{E}Z(t) = (2t/\pi)^{1/2} \quad \text{and} \quad \text{var}Z(t) = \left(1 - \frac{2}{\pi}\right)t. \quad \bullet$$

(16) **Example: Integrated Wiener Process** This is obtained from $W(t)$ by setting

$$V(t) = \int_0^t W(u) du.$$

Note that the integral exists because $W(t)$ is continuous. The process $V(t)$ has multivariate normal joint distributions because $W(t)$ does, but $V(t)$ is not in this case a Markov process. Clearly, $\mathbf{E}V(t) = 0$, and you can show that

$$\text{var}V(t) = \frac{1}{3}t^3$$

and

$$\text{cov}(V(s), V(t)) = \frac{1}{2}s^2(t - s/3), \quad s \leq t. \quad \bullet$$

We look at some even more interesting functions of the Wiener process shortly, but first we turn aside to make some important remarks. In our earlier work on Markov processes, the idea of the first passage time T of the process to some value was most important. This is for two main reasons. First, such first passage times are often naturally important in the real world that our processes describe. The second reason is that it is often useful to condition some event or expectation on the value of T , thus yielding immediate results or at least tractable equations for solution. These may be difference or differential or integral equations, or combinations of these. The key to the success of this approach is that the Markov property of the process is preserved at such first passage times, as we proved in the discrete case in Example 9.3.17.

It is crucial to our success in studying the Wiener process that these properties remain true; we summarize the basics here.

(17) **Definition** (a) The first passage time to b of the Wiener process $W(t)$ is given by

$$T_b = \inf\{t: W(t) = b\}.$$

(b) The first exit time from (a, b) of the Wiener process $W(t)$ is given by

$$T = \inf\{t: W(t) \notin (a, b)\}, \quad a < 0, b > 0.$$

Remark We may write $T = T_a \wedge T_b$, with an obvious notation. ▲

- (18) **Definition** Let S be a nonnegative random variable. If the event $S \leq t$ has probability 0 or 1, given $W(s)$ for $0 \leq s \leq t$, then S is called a *stopping time* (or *Markov time*) for $W(t)$. ▲

The two key properties (whose proofs we omit) are these:

- (19) **Lemma** The first passage times T_b and $T = T_a \wedge T_b$ are stopping times for $W(t)$.
- (20) **Theorem: Strong Markov Property** If S is a stopping time for the Wiener process $W(t)$, then $W(S+t) - W(S)$, $t \geq 0$, is a Wiener process independent of $W(s)$, $s \leq S$.

This last result says that the Markov property is preserved at Markov (or stopping) times, and this fact is extremely important and useful. Probabilists have been exceptionally inventive in exploiting the symmetry of the Wiener process coupled with the strong Markov property, but we have space for only some simple illustrative results about the maxima of the Wiener process and the Brownian Bridge.

- (21) **Example: Maximum** Let $M(t) = \max\{W(s), 0 \leq s \leq t\}$. We note these two useful points: first, for $c > 0$,

$$(22) \quad \{M(t) \geq c\} \supseteq \{W(t) \geq c\}.$$

Second, for $c > 0$, denoting the first passage time to c by T_c ,

$$(23) \quad \{M(t) \geq c\} \equiv \{T_c \leq t\},$$

and after T_c the process has a symmetric distribution about c that is independent of $W(t)$, $t \leq T_c$.

Therefore,

$$\begin{aligned} \mathbf{P}(W(t) \leq c | M(t) \geq c) &= \mathbf{P}(W(t) \geq c | M(t) \geq c) \\ &= \mathbf{P}(\{W(t) \geq c\} \cap \{M(t) \geq c\}) / \mathbf{P}(M(t) \geq c) \\ &= \mathbf{P}(W(t) \geq c) / \mathbf{P}(M(t) \geq c), \quad \text{using (22)}. \end{aligned}$$

Hence,

$$\begin{aligned} 1 &= \mathbf{P}(W(t) \leq c | M(t) \geq c) + \mathbf{P}(W(t) \geq c | M(t) \geq c) \\ &= 2\mathbf{P}(W(t) \geq c) / \mathbf{P}(M(t) \geq c) \end{aligned}$$

and so

$$\begin{aligned} \mathbf{P}(M(t) \geq c) &= 2\mathbf{P}(W(t) \geq c) \\ &= \mathbf{P}(|W(t)| \geq c) \\ &= \sqrt{\frac{2}{\pi t}} \int_c^\infty \exp\left(-\frac{1}{2} \frac{w^2}{t}\right) dw. \end{aligned}$$

It follows that M has the density

$$(24) \quad f_M(x) = \sqrt{\frac{2}{\pi t}} \exp\left(-\frac{x^2}{2t}\right), \quad x \geq 0. \quad \bullet$$

(25) **Corollary: First passage** Because $\{T_c \leq t\} \equiv \{M(t) \geq c\}$ it is immediate by setting $w = c(t/v)^{\frac{1}{2}}$ in the integral and differentiating for t that

$$f_{T_c}(t) = \frac{|c|}{(2\pi t^3)^{1/2}} \exp\{-\frac{c^2}{2t}\}, \quad t \geq 0. \quad \blacksquare$$

We can use much the same idea on the Brownian bridge.

(26) **Example** If $B(t)$ is a Brownian Bridge, $0 \leq t \leq 1$, show that for $c > 0$

$$\mathbf{P}(\max_{0 < t < 1} B(t) > c) = e^{-2c^2}.$$

Solution Once again, we use the fact that after T_c , the Wiener process is symmetrically distributed about c , and independent of the process before T_c . Hence,

$$\begin{aligned} \mathbf{P}(T_c < 1; 0 \leq W(1) < \varepsilon) &= \mathbf{P}(T_c < 1, 2c - \varepsilon < W(1) \leq 2c) \\ &= \mathbf{P}(2c - \varepsilon < W(1) \leq 2c) \\ &= \frac{1}{\sqrt{2\pi}} e^{-2c^2} \varepsilon + o(\varepsilon) \end{aligned}$$

But, by definition of $B(t)$, for $0 \leq t \leq 1$,

$$\begin{aligned} \mathbf{P}(\max B(t) > c) &= \lim_{\varepsilon \rightarrow 0} \mathbf{P}(T_c < 1 | 0 \leq W(1) \leq \varepsilon) \\ &= \lim_{\varepsilon \rightarrow 0} \mathbf{P}(T_c < 1; 0 \leq W(1) < \varepsilon) / \mathbf{P}(0 \leq W(1) \leq \varepsilon) \\ &= \lim_{\varepsilon \rightarrow 0} \left\{ \frac{1}{\sqrt{2\pi}} e^{-2c^2} \varepsilon + o(\varepsilon) \right\} / \{\phi(0)\varepsilon + o(\varepsilon)\} = e^{-2c^2}. \quad \bullet \end{aligned}$$

Remark The above technique, in which we combined the symmetry of the Wiener process with the Markov property at T_c , is called the *reflection principle*. This name is used because one can display a geometric visualization by taking the segment of a relevant path of the process after T_c , and reflecting it in the line $y = c$.

We proved this for the simple random walk in Lemma 5.6.19 and used it in the Hitting Time Theorem 5.6.17. You can envisage the same idea for the Wiener process by drawing a much spikier version of Figure 5.2. However, the proof in this case is beyond our scope.

Next, recall that earlier in this section we introduced several functions of the Wiener process, of various types. It turns out that functions of the Wiener process that are martingales turn out to be particularly useful.

(27) **Theorem** If $W(t)$ is a standard Wiener process, the following are martingales.

- (a) $W(t)$.
- (b) $Q(t) = \{W(t)\}^2 - t$.
- (c) $Z(t) = \exp(aW(t) - a^2t/2)$, $a \in \mathbb{R}$.

Proof Because $W(t) = W(s) + [W(t) - W(s)]$ and $W(t)$ has independent increments:

(a)

$$\begin{aligned}\mathbf{E}(W(t)|W(u), 0 \leq u \leq s) &= \mathbf{E}(W(s) + W(t) - W(s)|W(u), 0 \leq u \leq s) \\ &= W(s).\end{aligned}$$

(b) See Example (9.2.3).

(c)

$$\mathbf{E}\{\exp(aW(s) + aW(t) - aW(s))|W(u), 0 \leq u \leq s\} = e^{aW(s)} \exp\left(+\frac{1}{2}a^2(t-s)\right),$$

because $a(W(t) - W(s))$ is $N(0, a^2(t-s))$. The result follows.

For some applications, see Examples 9.22 and 9.23. ■

Remark There is a remarkable converse to this theorem. If $X(t)$ is a continuous process with $X(0) = 0$ such that $X(t)$ and $X^2(t) - t$ are both martingales, then $X(t)$ is the Wiener process. We omit the proof.

In fact, the existence of the third of the above martingales is particularly important. It asserts that

$$(28) \quad \mathbf{E} \exp\left(\theta W(t) - \frac{1}{2}\theta^2 t | W(u), 0 \leq u \leq s\right) = \exp\left(\theta W(s) - \frac{1}{2}\theta^2 s\right).$$

If we differentiate (28) with respect to θ and then set $\theta = 0$, we find that

$$\mathbf{E}(W(t)|W(u), 0 \leq u \leq s) = W(s),$$

which is simply the statement that $W(t)$ is a martingale. Differentiating (28) repeatedly with respect to θ , and then setting $\theta = 0$ at the appropriate moment, suggests that all the following (and many more) are martingales:

$$(29) \quad W(t)^2 - t$$

$$(30) \quad W(t)^3 - 3tW(t)$$

$$(31) \quad W(t)^4 - 6tW(t)^2 + 3t^2.$$

These can all be verified to be martingales just as we did in Theorem (27), but for the higher order functions of $W(t)$ it is a relief to know that the formal operation of differentiating with respect to θ and setting $\theta = 0$ can be proved to yield these martingales.

Just as we found in Chapter 8, such martingales are particularly useful when taken with a stopping time T . Because we are working with continuous martingales having continuous parameter, there are many technical considerations. We omit all the details and simply state the useful.

(32) **Theorem:** *Optional Stopping* Let $M(t)$ be a continuous martingale for $t \geq 0$. Suppose T is a stopping time for $M(t)$ such that $\mathbf{P}(T < \infty) = 1$. Then $M(t \wedge T)$ is also a martingale; further, $\mathbf{E}M(T) = \mathbf{E}M(0)$ if any of the following hold for some constant K :

- (a) $|M(t)| \leq K < \infty$, for $t \leq T$.
 (b) $\mathbf{E} \sup_t |M(T \wedge t)| \leq K < \infty$.
 (c) $\mathbf{E} M(T) \leq K < \infty$, and

$$\lim_{t \rightarrow \infty} \mathbf{E}(M(t)I(T > t)) = 0.$$

- (d) $T \leq K < \infty$.

We give several applications later; here is one simple basic corollary.

(33) **Example: Exiting a Strip** Let T be the time at which $W(t)$ first hits a or b , where $a < 0$ and $b > 0$. Show that

$$\mathbf{P}(W(T) = a) = \frac{b}{b - a}, \quad \mathbf{P}(W(T) = b) = -\frac{a}{b - a}.$$

Solution We have that $W(t)$ is a martingale, and $|W(t)| \leq a \vee b = \max\{a, b\}$. It is easy to see that $\mathbf{P}(T < \infty) = 1$, and so by the first form of the optional stopping theorem

$$(34) \quad 0 = \mathbf{E}W(T) = a\mathbf{P}(W(T) = a) + b\mathbf{P}(W(T) = b).$$

Now using $\mathbf{P}(T < \infty) = 1$, we have

$$(35) \quad 1 = \mathbf{P}(W(T) = a) + \mathbf{P}(W(T) = b),$$

and solving (34) and (35) gives the result. ●

We conclude this section with a brief look at another important application of martingales and the Wiener process.

(36) **Example: The Option-Pricing Martingale** A popular model for a simple market comprises two available assets: a bond whose value $B(t)$ grows at a continuously compounded constant interest rate r , so that $B(t) = B(0)e^{rt}$, and a stock whose price per unit is some suitable random process $S(t)$. The model assumes that no dividends taxes or commissions are paid; you may buy negative amounts of stock, which is called “selling short” and may lead to a “short position.”

In 1900, Bachelier suggested the Wiener process as a model for $S(t)$, but this has the drawback that it may be negative, which stock prices never are. More recently, it has been standard to assume that $S(t)$ follows a geometric Brownian motion, which is to say that for some constants μ and σ^2 ,

$$S(t) = \exp\{(\mu t + \sigma W(t))\},$$

where $W(t)$ is the standard Wiener process.

In many practical situations, to reduce uncertainty and risk, it is desirable to acquire an option to purchase stock at some time T in the future. One simple and popular type is the European call option. This confers the right (but not the obligation) to buy a unit of stock at time T (the exercise date) at cost K (the strike price). Clearly, the value of this option at time T is

$$V(T) = (S(T) - K)^+ = \max\{S(T) - K, 0\}.$$

The key question is, what is the fair price $V(t)$ for this option at any other time t , where $0 \leq t < T$?

An answer to this question is suggested by our interpretation of a martingale as a fair game. The point is that because you can always invest in the bond with interest rate r , the future value of the stock at time t must be discounted now by e^{-rt} .

That is to say, suppose at time 0 you intend to buy a unit of stock at time s and sell it at time $t > s$. The present discounted value of the purchase is $e^{-rs}S(s)$, the present discounted value of the sale is $e^{-rt}S(t)$. If this is to be a “fair game,” then the expected value of the sale should equal the purchase price so that the expectation E_0 taken with respect to the pay-off odds of this “fair game” must satisfy

$$(37) \quad E_0(e^{-rt}S(t)|S(u); 0 \leq u \leq s) = e^{-rs}S(s).$$

That is to say, $e^{-rt}S(t)$ is a martingale. It turns out that if we set $\mu = r - \sigma^2/2$, then $e^{-rt}S(t)$ is indeed a martingale. It can further be shown also that $e^{-rt}V(t)$ is a martingale with respect to these same pay-off odds, and then it follows that the fair price of the option at $t = 0$ is

$$(38) \quad v = E_0[e^{-rT}(S(T) - K)^+],$$

where the expectation E_0 is taken according to the pay-off odds fixed in (37).

See Example 9.26 for more details. Note that the heuristic argument above can indeed be made rigorous, but the details are well beyond our scope here. ●

It sometimes happens that students seeing the valuation (38) for the first time find it a little counterintuitive, because the expected value on the right side is taken with respect to the pay-off odds satisfying (37), not with respect to the “real” probability distribution of the process $S(t)$.

The following analogy, metaphor, or parable is often useful in reorienting such misdirected intuition.

(39) **Example: The Bookmaker** Suppose a bookie is setting the pay-off odds (making a book) for a two-horse race. As it happens, these horses (A and B) are identical twins, their record in 1000 head-to-head races is 500 wins each, and they have just run shoulder to shoulder in training. *A priori*, the probability of A winning is $\mathbf{P}(A) = \frac{1}{2}$. But the market (the gamblers) has laid \$5000 on A to win and \$10,000 on B to win.

If the bookie sets the pay-off odds to be evens (the same as the *a priori* probabilities), then she loses \$5000 if B wins. Of course, she gains \$5000 if A wins, and the *a priori* expectation of her outcome is $\frac{1}{2}5000 - \frac{1}{2}5000 = 0$, which is “fair.” But if the bookie wanted to take risks, she could simply gamble, and this is not why she is a bookie.

In fact, suppose she sets the pay-off odds on A to be $2 : 1$, and those on B to be $1 : 2$. Then whatever the outcome of the race, she is all square, which makes it a “fair game.” These pay-off odds correspond to a probability distribution

$$P_0(A) = \frac{1}{3}; \quad P_0(B) = \frac{2}{3},$$

which is different from the empirical *a priori* distribution. (In practice, of course, she would shorten all the pay-off odds to guarantee a profit or arbitrage.)

But the point for the price of the option is clear; it is determined by the fair pay-off odds arising from actual market opportunities, not from any theoretical *a priori* valuation of the stocks. ●

9.10 Review and Checklist for Chapter 9

The intuitive idea of a Markov process is that conditional on its present state, the future is independent of the past. We make this idea precise for Markov processes in discrete and continuous time. For those with discrete state space, we derive the Chapman–Kolmogorov equations and use them to examine the evolution of the chain over time. We consider first passage times and recurrence times, and examine the link to stationary distributions. Then we discuss the link between stationary distributions and the possible long-run behaviour of the chain. We consider Poisson and birth processes, in particular. Finally we turn to Markov processes, with a continuous state space, that do not have jump transitions. The first and classic example of such a process is the Wiener process model for Brownian motion. We examine many of its properties, with a brief glance at other diffusion processes derived from the Wiener process. Martingales derived from the Wiener process turn out to be especially useful.

Synopsis of Notation and Formulae

For a discrete time-homogeneous Markov chain X_n , we have

The transition probabilities: $p_{ij} = \mathbf{P}(X_{n+1} = j | X_n = i)$

The m -step transition probabilities: $p_{ij}(m) = \mathbf{P}(X_{n+m} = j | X_n = i)$

The Chapman–Kolmogorov equations:

$$p_{ik}(m+n) = \sum_{j \in S} p_{ij}(m)p_{jk}(n).$$

The first passage time from i to $k \neq i$ is T_{ik} , with mean first passage time $\mu_{ik} = \mathbf{E}T_{ik}$.

The recurrence time of i is $T_i = \min\{n > 0 : X_n = i | X_0 = i\}$, with mean recurrence time $\mu_i = \mathbf{E}T_i$. If $\mu_i < \infty$, then i is nonnull.

A stationary measure is a nonnegative solution $(x_i; i \in S)$ of

$$x_k = \sum_{i \in S} x_i p_{ik}.$$

A stationary distribution is a stationary measure x_i such that $\sum_{i \in S} x_i = 1$.

In this case, we write $x_i = \pi_i$.

A chain is regular if $p_{ij}(n_0) > 0$ for all i, j , and some $n_0 < \infty$. For a finite state space regular chain,

- There is a unique stationary distribution π .
- $\pi_i \mu_i = 1$ for all $i \in S$.
- For all pairs i, j , as $n \rightarrow \infty$, $p_{ij}(n) \rightarrow \pi_j = \mu_j^{-1}$.

If the chain is not regular or has infinite state space, a wider range of behaviour is possible.

For a Markov chain X_t with continuous parameter, we have

The transition probabilities: $p_{ij}(t) = \mathbf{P}(X_{s+t} = j | X_s = i)$.

The Chapman–Kolmogorov equations:

$$p_{ik}(s+t) = \sum_{j \in S} p_{ij}(s) p_{jk}(t).$$

The first passage time from i to $k \neq i$ is

$$T_{ik} = \inf \{t \geq 0 : X_t = k | X_0 = i\},$$

with mean first passage time $\mu_{ik} = \mathbf{E}T_{ik}$.

A stationary measure is a nonnegative solution $(x_i; i \in S)$ of

$$x_k = \sum_i x_i p_{ik}(t).$$

A stationary distribution is a stationary measure such that $\sum_i x_i = 1$. In this case, we write $x_i = \pi_i$.

For a finite state space chain with $p_{ij}(t) > 0$ for all i, j ,

- There is a unique stationary distribution π .
- For all pairs i, j as $t \rightarrow \infty$, $p_{ij}(t) \rightarrow \pi_j$.

If the chain is not irreducible or has infinite state space, a much wider range of behaviour is possible.

Checklist of Terms

9.1 Markov property

Markov chain
imbedding

9.2 transition matrix

doubly stochastic matrix
Chapman–Kolmogorov
 n -step transition probabilities
regular chain
irreducible
aperiodic
absorbing state

9.3 mean first passage time, mean recurrence time

Markov property at first passage times

- 9.4 Stationary distribution
- 9.5 limiting distribution
 - coupling
 - transient
 - persistent
 - null
- 9.6 Chapman–Kolmogorov equations
- 9.7 Poisson process
 - birth process
- 9.8 forward equations
 - stationary distribution
- 9.9 Wiener process
 - Brownian bridge
 - Ornstein–Uhlenbeck process
 - drift
 - reflection
 - geometric Wiener process
 - martingales
 - optional stopping
 - option pricing
 - strong Markov property

WORKED EXAMPLES AND EXERCISES

9.11 Example: Crossing a Cube

One vertex O of a unit cube is at the origin $(0, 0, 0)$. The others are at $(0, 0, 1)$, $(0, 1, 0)$ and so on. A particle performs a random walk on the vertices of this cube as follows. Steps are of unit length, and from any vertex it steps in the x direction with probability α , the y direction with probability β or the z direction with probability γ , where $\alpha + \beta + \gamma = 1$.

- (a) Let T be the first passage time from O to V . Find $\mathbf{E}(s^T)$, and deduce that

$$\mathbf{E}(T) = 1 + \frac{1}{\alpha} + \frac{1}{\beta} + \frac{1}{\gamma}.$$

- (b) Let X be the number of visits that the walk makes to V before the first return to O . Show that $\mathbf{E}(X) = 1$.

Solution The walk visits O whenever an even number of steps has been taken in all three possible directions (x , y , and z directions). Thus,

$$u(2n) = \mathbf{P}(\text{the walk visits } O \text{ at the } 2n\text{th step}) = \sum_{i+j+k=n} \alpha^{2i} \beta^{2j} \gamma^{2k} \frac{(2n)!}{(2i)!(2j)!(2k)!}.$$

Now write $\beta + \gamma - \alpha = a$, $\alpha + \gamma - \beta = b$, and $\alpha + \beta - \gamma = c$. It is easy to check, by expanding each side, that

$$\begin{aligned} 4u(2n) &= (\alpha + \beta + \gamma)^{2n} + (\beta + \gamma - \alpha)^{2n} + (\alpha + \gamma - \beta)^{2n} + (\alpha + \beta - \gamma)^{2n} \\ &= 1 + a^{2n} + b^{2n} + c^{2n}. \end{aligned}$$

Hence,

$$U(s) = \sum_{n=0}^{\infty} s^{2n} u(2n) = \frac{1}{4} \left(\frac{1}{1-s^2} + \frac{1}{1-a^2s^2} + \frac{1}{1-b^2s^2} + \frac{1}{1-c^2s^2} \right).$$

Similarly, starting from O , the walk visits V whenever an odd number of steps has been taken in all three possible directions. Hence,

$$\begin{aligned} u_V(2n+1) &= \mathbf{P}(\text{the walk visits } V \text{ at the } (2n+1)\text{th step}) \\ &= \sum_{i+j+k=n-1} \alpha^{2i+1} \beta^{2j+1} \gamma^{2k+1} \frac{(2n+1)!}{(2i+1)!(2j+1)!(2k+1)!}. \end{aligned}$$

Now it is easy to check as above that $4u_V(2n+1) = 1 - a^{2n+1} - b^{2n+1} - c^{2n+1}$.

Hence,

$$U_V(s) = \sum_{n=1}^{\infty} u_V(2n+1) s^{2n+1} = \frac{s^3}{4} \left(\frac{1}{1-s^2} - \frac{a^3}{1-a^2s^2} - \frac{b^3}{1-b^2s^2} - \frac{c^3}{1-c^2s^2} \right).$$

Now we use (9.3.21) to see that

$$\mathbf{E}(s^T) = \frac{U_V(s)}{U(s)}.$$

Hence, evaluating $\frac{d}{ds} \mathbf{E}(s^T)$ at $s = 1$ yields $\mathbf{E}(T) = 1 + \alpha^{-1} + \beta^{-1} + \gamma^{-1}$, as required.

(b) Let ρ be the probability that the walk returns to O before ever visiting V . By symmetry, this is also the probability that a walk started at V never visits O before returning to V . For $X = k$, it is necessary for the walk to reach V before revisiting O , then revisit V on $k-1$ occasions without visiting O , and finally return to O with no further visit to V . Hence, $\mathbf{P}(X = k) = (1 - \rho)\rho^{k-1}(1 - \rho)$, by the Markov property. Therefore,

$$\mathbf{E}(X) = \sum_{k=1}^{\infty} k\rho^{k-1}(1 - \rho)^2 = 1.$$

- (1) **Exercise** Suppose that at every step the walk may remain at its current vertex with probability δ , where now $\alpha + \beta + \gamma + \delta = 1$. Find:
- (a) The mean recurrence time of O ; (b) $\mathbf{E}(X)$; (c) $\mathbf{E}(T)$.
- (2) **Exercise** Let W be the vertex $(1, 1, 0)$, and define \hat{T} to be the number of steps until the walk first visits V or W starting from O . (That is to say, \hat{T} is the first passage time from O to $\{V, W\}$.) Show that $\mathbf{E}(\hat{T}) = (\alpha^{-1} + \beta^{-1})\gamma$.

9.12 Example: Reversible Chains

A collection of random variables $(X(n); n \in \mathbb{Z})$ is called *reversible* if $(X(n_1), X(n_2), \dots, X(n_r))$ has the same distribution as $(X(m - n_1), X(m - n_2), \dots, X(m - n_r))$ for all m and $n_1 < n_2 < \dots < n_r$. Let X be an irreducible nonnull recurrent aperiodic Markov chain.

(a) Prove that the Markov chain X with transition probabilities p_{ij} is reversible if and only if it is stationary and there exist $(\pi_i; i \in S)$ such that for all $i, j \in S$

$$(1) \quad \pi_i > 0,$$

$$(2) \quad \sum_{i \in S} \pi_i = 1,$$

$$(3) \quad \pi_i p_{ij} = \pi_j p_{ji}.$$

(b) Prove further that if X is stationary then it is reversible if and only if

$$(4) \quad p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_r i_1} = p_{i_1 i_r} p_{i_r i_{r-1}} \cdots p_{i_2 i_1}$$

for any finite sequence of states i_1, i_2, \dots, i_r in S .

Solution (a) The truth of (1), (2), and (3) implies that π_i is the stationary distribution of X , for summing (3) over j yields $\sum_i \pi_i p_{ij} = \pi_j$, as required.

Next we note that (3) implies

$$(5) \quad \pi_i p_{ij}(n) = \pi_j p_{ji}(n).$$

To see this, consider any n -step path from i to j , and then using (3) gives

$$\pi_i p_{ii_1} p_{i_1 i_2} \cdots p_{i_{n-1} j} = \pi_i \pi_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} j} = p_{i_1 i} p_{i_2 i_1} \cdots p_{j i_{n-1}} \pi_j$$

after repeated applications of (3). Now summing over all paths from i to j gives (5).

Now applying (5) shows that

$$\begin{aligned} \mathbf{P}(X_{n_1} = i_1, X_{n_2} = i_2, \dots, X_{n_r} = i_r) &= \pi_{i_1} p_{i_1 i_2}(n_2 - n_1) \cdots p_{i_{r-1} i_r}(n_r - n_{r-1}) \\ &= \mathbf{P}(X_{m-n_1} = i_1, \dots, X_{m-n_r} = i_r), \end{aligned}$$

as required. The converse is obvious, and so we have finished (a).

(b) If the chain is reversible then (3) holds. Hence,

$$p_{i_1 i_2} \cdots p_{i_r i_1} = \frac{\pi_{i_2}}{\pi_{i_1}} p_{i_2 i_1} \cdots \frac{\pi_{i_1}}{\pi_{i_r}} p_{i_r i_1},$$

which is (4) because all the π_i cancel successively. Conversely, if (4) holds, we may choose two states i_1 and i_r (say) as fixed and equal, so $i_1 = i_r = j$. Then summing (4) over the remaining indices yields

$$p_{ij}(r-1)p_{ji} = p_{ji}(r-1)p_{ij}.$$

Now, as the chain is aperiodic and nonnull recurrent, we can let $r \rightarrow \infty$ and obtain (3), as required.

Remark Equations (3) are called the *detailed balance equations*. This example is important because a remarkable number of Markov chains are reversible in equilibrium (particularly those encountered in examinations). In the exercises, you may assume that X is aperiodic and irreducible.

- (6) **Exercise** Let $(X_n; n \in \mathbb{Z})$ be a Markov chain, and let $Y_n = X_{-n}$ be the chain X reversed. Show that for $n_1 < n_2 < \dots < n_r$

$$\mathbf{P}(Y_{n_r} = k | Y_{n_1}, \dots, Y_{n_{r-1}}) = \mathbf{P}(Y_{n_r} = k | Y_{n_{r-1}}).$$

Show that if X has stationary distribution π and transition probabilities \mathbf{P} , then, in equilibrium, Y is a Markov chain with transition probabilities $q_{ij} = \pi_j \pi_i^{-1} p_{ji}$.

- (7) **Exercise 6 Continued** Let X be a Markov chain, with transition probabilities p_{ij} . Show that if there exists a stochastic matrix (q_{ij}) and a mass function (π_i) such that $\pi_i q_{ij} = \pi_j p_{ji}$, then q_{ij} is the transition matrix of X reversed and π_i is its stationary distribution.
- (8) **Exercise** Let X_n be a Markov chain on the nonnegative integers with transition matrix

$$p_{ij} = \begin{cases} \lambda_i & \text{if } j = i + 1 > 0 \\ \mu_i & \text{if } j = i - 1 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$p_{00} = \mu_0.$$

Show that X is reversible in equilibrium. Deduce that X has a stationary distribution if and only if $\sum_{n=1}^{\infty} \prod_{k=1}^n \frac{\lambda_{k-1}}{\mu_k} < \infty$.

- (9) **Exercise** Let X_n and Y_n be independent reversible Markov chains with stationary distributions π and ν , respectively. Let $Z_n = (X_n, Y_n)$. Show that Z_n is reversible in equilibrium.

9.13 Example: Diffusion Models

- (a) Two separate containers together contain m distinguishable particles. At integer times, $t = 1, 2, \dots$ one of the particles is selected at random and transferred to the other container. Let X_n be the number of particles in the first container after the n th transfer.
- (i) Show that $(X_n; n \geq 0)$ is a Markov chain and write down its transition matrix.
- (ii) Find the stationary distribution of X .
- (b) The two containers C_1 and C_2 are now separated by a semipermeable membrane. At each time t , a particle is selected at random; if it is in C_1 , then it is transferred with probability α ; if it is in C_2 , then it is transferred with probability β . Otherwise, the particles remain where they are. Show that X has stationary distribution

$$(1) \quad \pi_i = \frac{\alpha^{m-i} \beta^i}{(\alpha + \beta)^m} \binom{m}{i}.$$

Solution (a) Given $X_0, \dots, X_n = j$, the probability that a particle in C_1 is selected for transfer is j/m , and the probability that a particle in C_2 is selected for transfer is $(m - j)/m$. Hence, X is a Markov chain and

$$p_{j,j+1} = (m - j)/m,$$

$$p_{j,j-1} = j/m,$$

$$p_{jk} = 0 \quad \text{when } |k - j| \neq 1.$$

The chain is of the type considered in Exercise 9.12.8. Therefore, the chain is reversible and the stationary distribution may be obtained by solving $\pi_i p_{ij} = \pi_j p_{ji}$. Thus,

$$\pi_{i+1} = \pi_i p_{i,i+1}/p_{i+1,i} = \pi_i(m-i)/(i+1) = \pi_0 \binom{m}{i+1} \quad \text{on iterating.}$$

Because $\sum_0^m \pi_i = 1$, it follows that

$$\pi_0^{-1} = \sum_{i=0}^m \binom{m}{i} = 2^m$$

and so $\pi_i = \binom{m}{i} 2^{-m}$. This is a symmetric binomial distribution.

(b) By the same reasoning as given in (a), this is a Markov chain with

$$\begin{aligned} p_{i,i+1} &= \beta(m-i)/m \\ p_{i,i-1} &= \alpha i/m \\ p_{ii} &= 1 - \alpha i/m - \beta(m-i)/m \\ p_{ij} &= 0 \quad \text{when } |i-j| > 1. \end{aligned}$$

Again, this is reversible, so the stationary distribution is given by

$$\begin{aligned} (2) \quad \pi_{i+1} &= \pi_i p_{i,i+1}/p_{i+1,i} = \pi_i \beta(m-i)/(\alpha(i+1)) \\ (3) \quad &= \pi_0 \left(\frac{\beta}{\alpha}\right)^{i+1} \binom{m}{i+1} \quad \text{on iterating.} \end{aligned}$$

Now requiring $\sum \pi_i = 1$ yields

$$\pi_i = \frac{\alpha^{m-i} \beta^i}{(\alpha + \beta)^m} \binom{m}{i},$$

which is an asymmetric binomial distribution.

Remark This is the Ehrenfest model for diffusion (and heat transfer) (1907).

- (4) **Exercise** Is it true that as $n \rightarrow \infty$, $p_{ij}(n) \rightarrow \pi_j$ in either case (a) or case (b)?
- (5) **Exercise** (a) Show that $\mathbf{E}(X_n) = (1 - \frac{2}{m})^n (\mathbf{E}(X_0) - \frac{m}{2}) + \frac{m}{2}$, when $\alpha = \beta = 1$.
(b) What is $\mathbf{E}(X_n)$ when $\alpha \neq 1 \neq \beta$?
- (6) **Exercise: Bernoulli Diffusion Model** Two adjacent containers C_1 and C_2 each contain m particles. Of these $2m$ particles, m are of type A and m are of type B. At $t = 1, 2, \dots$ one particle is selected at random from each container, and these two particles are each transferred to the other container. Let X_n be the number of type A particles in C_1 .
(a) Show that X is a Markov chain and write down its transition matrix.
(b) Find the stationary distribution.
(c) Is the stationary distribution the limiting distribution of X ?
- (7) **Exercise** In the above exercise, suppose that the containers are separated by a semipermeable membrane. In this case, if particles of different types are chosen they are exchanged with probability α if the type A particle is in C_1 , or with probability β if the type A particle is in C_2 . Find the stationary distribution of X .

9.14 Example: The Renewal Chains

Let $(f_n; n \geq 1)$ satisfy $f_n \geq 0$ and

$$(1) \quad \sum_{i=1}^n f_i \leq 1.$$

Define a sequence $(u_n; n \geq 0)$ by $u_0 = 1$ and

$$(2) \quad u_n = \sum_{r=1}^n f_r u_{n-r}; \quad n \geq 1.$$

Such a sequence $(u_n; n \geq 0)$ is called a *renewal sequence*.

(a) Show that $(u_n; n \geq 0)$ is a renewal sequence if and only if there is a Markov chain U_n such that for some state $s \in S$,

$$u_n = \mathbf{P}(U_n = s | U_0 = s).$$

(b) Let X be a random variable having the probability mass function $(f_n; n \geq 1)$. Show that the chain U is recurrent if $\sum_n f_n = 1$ and nonnull if $\mathbf{E}(X) < \infty$.

(c) Explain why u_n is called a renewal sequence.

Solution (a) Let u_n and f_n be as defined above. Define the sequence $(F_n; n \geq 0)$ by $F_0 = 0$ and

$$F_n = \sum_{r=1}^n f_r; \quad n \geq 1.$$

Next, let $(U_n; n \geq 0)$ be a Markov chain taking values in the nonnegative integers, with $U_0 = 0$, and having transition probabilities

$$(3) \quad p_{i0} = \frac{f_{i+1}}{1 - F_i}$$

$$(4) \quad p_{i,i+1} = 1 - p_{i0} = \frac{1 - F_{i+1}}{1 - F_i}; \quad i \geq 0.$$

Now let us calculate the first return probability $f_{00}(n)$, that is the probability that the chain first returns to 0 at the n th step. At each stage, the chain either does so return or increases by 1. Hence,

$$(5) \quad \begin{aligned} f_{00}(n) &= p_{01} p_{12} \cdots p_{n-2,n-1} p_{n-1,0} \\ &= \frac{1 - F_1}{1 - F_0} \cdot \frac{1 - F_2}{1 - F_1} \cdots \frac{1 - F_{n-1}}{1 - F_{n-2}} \cdot \frac{f_n}{1 - F_{n-1}} = f_n. \end{aligned}$$

It follows by conditional probability and the Markov property that

$$(6) \quad \mathbf{P}(U_n = 0) = \sum_{r=1}^n f_{00}(r) \mathbf{P}(U_{n-r} = 0) = \sum_{r=1}^n f_r \mathbf{P}(U_{n-r} = 0).$$

Because $\mathbf{P}(U_0 = 0) = 1$, it follows by comparison with (2) that $u_n = \mathbf{P}(U_n = 0 | U_0 = 0)$, as required.

(b) First, observe that if $F_j = 1$ for some $j < \infty$, then the chain is finite. Obviously, in this case, $\mathbf{E}(X) < \infty$ and the chain is recurrent and nonnull. Otherwise, if $\sum f_i = 1$, then $\sum_n f_{00}(n) = 1$ by (5), and hence the chain is recurrent. Because it is irreducible we can settle whether it is nonnull by seeking a stationary distribution π . Any such stationary distribution satisfies

$$\begin{aligned} \pi_{i+1} &= \pi_i p_{i,i+1}; \quad i \geq 0 \\ &= \pi_i \frac{1 - F_{i+1}}{1 - F_i} = \pi_0 (1 - F_{i+1}) \quad \text{on iterating.} \end{aligned}$$

Hence, π is a stationary distribution satisfying $\sum_i \pi_i = 1$ if $\sum_i (1 - F_i) < \infty$. That is to say, U is nonnull if $\mathbf{E}(X) < \infty$, and then

$$(7) \quad \pi_i = \frac{1 - F_i}{\mathbf{E}(X)} = \frac{\mathbf{P}(X > i)}{\mathbf{E}(X)}.$$

(c) Recall the recurrent event or renewal processes defined in Section 6.7. Events may occur at integer times and the intervals between successive events are independent and identically distributed random variables $(X_i; i \geq 1)$, where $f_X(r) = f_r$. An event occurs at $n = 0$. Now the construction of the chain U in (3) and (4) allows us to identify visits of U to zero with the occurrence of an event in this renewal process. At any time n , the state U_n of the chain is just the time elapsed since the most recent event of the renewal process. Thus, U_n is the current life (or age) of the renewal process at time n .

Finally, u_n is just the probability that an event of the renewal process occurs at time n .

- (8) **Exercise** Show that if u_n and v_n are renewal sequences, then $u_n v_n$ is a renewal sequence.
- (9) **Exercise** Show that if u_n is a renewal sequence, then $(u_{nd}; u \geq 0)$ is a renewal sequence for any fixed d .
- (10) **Exercise** Let $(X_i; i \geq 1)$ be the interevent times of a discrete renewal process, and at any time n let B_n be the time until the next following event of the process. (That is, B_n is the excess life or balance of life.) Show that B_n is a Markov chain, and find the stationary distribution when it exists.
- (11) **Exercise** Write down the transition probabilities of the chain U_n reversed in equilibrium, and also write down the transition probabilities of B_n reversed in equilibrium. Explain your answers. (Hint: Use Exercise 9.11.6.)
- (12) **Exercise** Use Theorem 9.6.5 to show that $\lim_{n \rightarrow \infty} u_n = \pi_0$.
- (13) **Exercise** Recall the recurrent event (renewal) processes of Section 6.7. Use (12) to show that as $n \rightarrow \infty$, $u_n \rightarrow \frac{1}{\mathbf{E}(X_2)}$ and $v_n \rightarrow \frac{1}{\mathbf{E}(X_2)}$.

9.15 Example: Persistence

Let i and j be states of the Markov chain X . We write $i \rightarrow j$ if there exists $n < \infty$ such that $p_{ij}(n) > 0$. Let V_{ij} be the number of occasions on which X visits j , given that initially $X_0 = i$. Define

$$\eta_{ij} = \mathbf{P}(V_{ij} = \infty),$$

and recall that T_{ij} is the first passage time from i to j .

(a) Show that

$$\eta_{ii} = \begin{cases} 1 & \text{if } i \text{ is persistent} \\ 0 & \text{if } i \text{ is transient.} \end{cases}$$

(b) Show that

$$\eta_{ij} = \begin{cases} \mathbf{P}(T_{ij} < \infty) & \text{if } j \text{ is persistent} \\ 0 & \text{if } j \text{ is transient.} \end{cases}$$

(c) Show that if $i \rightarrow j$ and i is persistent then $\eta_{ij} = \eta_{ji} = 1$.

Solution (a) First note that $V_{ii} \geq 1$ if and only if $T_i < \infty$, so $\mathbf{P}(V_{ii} \geq 1) = \mathbf{P}(T_i < \infty)$. Next,

$$\mathbf{P}(V_{ii} \geq 2) = \mathbf{P}(V_{ii} \geq 2 | V_{ii} \geq 1) \mathbf{P}(V_{ii} \geq 1) = \mathbf{P}(V_{ii} \geq 2 | T_i < \infty) \mathbf{P}(V_{ii} \geq 1).$$

However, we have shown in Example 9.3.17 that the Markov property is preserved at T_i when it is finite, and therefore $\mathbf{P}(V_{ii} \geq 2 | T_i < \infty) = \mathbf{P}(V_{ii} \geq 1)$.

Now an obvious induction shows that

$$\mathbf{P}(V_{ii} \geq n) = (\mathbf{P}(V_{ii} \geq 1))^n = (\mathbf{P}(T_i < \infty))^n \rightarrow \begin{cases} 1 & \text{if } \mathbf{P}(T_i < \infty) = 1 \\ 0 & \text{if } \mathbf{P}(T_i < \infty) < 1 \end{cases},$$

as $n \rightarrow \infty$, as required.

(b) Using the same idea as in (a), we write

$$\begin{aligned} \mathbf{P}(V_{ij} \geq m) &= \mathbf{P}(V_{ij} \geq m | V_{ij} \geq 1) \mathbf{P}(V_{ij} \geq 1) \\ &= \mathbf{P}(V_{ij} \geq m | T_{ij} < \infty) \mathbf{P}(T_{ij} < \infty) = \mathbf{P}(V_{jj} \geq m - 1) \mathbf{P}(T_{ij} < \infty) \end{aligned}$$

because the Markov property is preserved at T_{ij} . Now allowing $m \rightarrow \infty$ gives the result, by (a).

(c) Because i is persistent

$$\begin{aligned} 1 &= \mathbf{P}(V_{ii} = \infty) = \mathbf{P}(\{V_{ij} = 0\} \cap \{V_{ii} = \infty\}) + \mathbf{P}(\{V_{ij} > 0\} \cap \{V_{ii} = \infty\}) \\ &\leq \mathbf{P}(V_{ij} = 0) + \mathbf{P}(\{T_{ij} < \infty\} \cap \{V_{ii} = \infty\}) \\ &= 1 - \mathbf{P}(T_{ij} < \infty) + \mathbf{P}(T_{ij} < \infty) \mathbf{P}(V_{ji} = \infty) \end{aligned}$$

because the Markov property is preserved at T_{ij} .

Hence, $\mathbf{P}(V_{ji} = \infty) \geq 1$, and therefore $\eta_{ji} = 1$. Hence, $\mathbf{P}(T_{ij} < \infty) = 1$ and so $j \rightarrow i$. It follows that $\eta_{ij} = 1$.

- (1) **Exercise** Show that if i is persistent and $i \rightarrow j$, then j is persistent.
- (2) **Exercise** Show that $\eta_{ij} = 1$ if and only if $\mathbf{P}(T_{ij} < \infty) = \mathbf{P}(T_j < \infty) = 1$.
- (3) **Exercise** Show that if $i \rightarrow j$ and $j \rightarrow i$, then i and j have the same class and period.
- (4) **Exercise** Show that if X is irreducible and persistent, then $\mathbf{P}(T_{ij} < \infty) = \mathbf{P}(T_{ji} < \infty) = 1$.
- (5) **Exercise** Show that if $i \rightarrow j$ but $j \not\rightarrow i$, then i is transient.

9.16 Example: First Passages and Bernoulli Patterns

Let X be a finite Markov chain with n -step transition probabilities $p_{ij}(n)$. As usual, T_{ij} is the first passage time from i to j , with mean μ_{ij} ; T_j is the recurrence time of j with mean μ_j . We write $F_{ij}(s) = \mathbf{E}(s^{T_{ij}})$, and $P_{ij}(s) = \sum_{n=0}^{\infty} p_{ij}(n)s^n$.

(a) A biased coin is tossed repeatedly, let X_n be the outcome of tosses $n + 1, n + 2,$ and $n + 3$; for $n \geq 0$. [Thus, $(X_n; n \geq 0)$ is a Markov chain with state space S comprising all triples using H and T , namely, $HHH, HTH, HHT,$ and so on.] Show that for any i and $j \in S$,

$$(1) \quad \mu_{ij} = [1 + p_{jj}(1) + p_{jj}(2) - p_{ij}(1) - p_{ij}(2)]\mu_j.$$

Deduce that if $i = HHH, j = THT$, and the coin is fair, then $\mu_{ij} = 10$ and $\mu_{ji} = 14$.

(b) Let D be a given subset of the states of a finite Markov chain $(Y_n; n \geq 0)$, and let T_{sD} be the first passage time from the state $s \notin D$ into D , with mean $\mu_{sD} = \mathbf{E}(T_{sD})$. Also, let ϕ_{sj} be the probability that the chain first enters D at the state j . Show that for $i \in D$

$$(2) \quad \mu_{si} = \mu_{sD} + \sum_{j \in D} \phi_{sj}\mu_{ji}.$$

(c) Hence, show that in an unlimited sequence of tosses of a fair coin, the probability that the consecutive sequence THT occurs before HHH is $\frac{7}{12}$.

Solution (a) From Theorem 6.2.13, we have that

$$(3) \quad \mu_{ij} = \lim_{s \uparrow 1} \frac{1 - F_{ij}(s)}{1 - s} = \lim_{s \uparrow 1} \frac{P_{jj}(s) - P_{ij}(s)}{(1 - s)P_{jj}(s)} \quad \text{by Theorem 9.3.20.}$$

Now because tosses of the coin are independent, we have

$$(4) \quad p_{ij}(n) = p_{jj}(n) \quad \text{for } n \geq 3.$$

Also, using Theorems 6.2.13 and 9.3.20 again gives

$$(5) \quad \lim_{s \uparrow 1} (1 - s)P_{jj}(s) = \mu_j^{-1},$$

and now (3), (4), and (5) give (1).

If the chance of a head is p , then $p_{ij}(1) = p_{ij}(2) = p_{jj}(1) = 0$ and $p_{jj}(2) = p(1 - p)$. Hence,

$$\mu_{ij} = [1 + p(1 - p)]\mu_j.$$

When the coin is fair $p = \frac{1}{2}$ and $\mu_j = 8$, so $\mu_{ij} = 10$. Likewise, $p_{ji}(1) = p_{ji}(2) = 0$ and $p_{ii}(1) = p, p_{ii}(2) = p^2$. Hence, when the coin is fair,

$$\mu_{ji} = \left(1 + \frac{1}{2} + \frac{1}{4}\right)8 = 14.$$

(b) Let D_j denote the event that the chain first enters D at j . Then

$$(6) \quad \begin{aligned} \mu_{si} &= \mathbf{E}(T_{si}) = \mathbf{E}(T_{si} - T_{sD} + T_{sD}) = \mathbf{E}(T_{si} - T_{sD}) + \mu_{sD} \\ &= \sum_{j \in D} \mathbf{E}(T_{si} - T_{sj} | D_j) \phi_{sj} + \mu_{sD}. \end{aligned}$$

However, given D_j , the chain continues its journey to i independently of the past, that is to say,

$$(7) \quad \mathbf{E}(T_{si} - T_{sj} | D_j) = \mu_{ji}.$$

Substituting (7) into (6) gives (2).

(c) Now let the outcomes of successive tosses be $(S_n; n \geq 1)$ and define the Markov chain $(Y_n; n \geq 0)$ by

$$\begin{aligned} Y_0 &= \phi = s \\ Y_1 &= S_1 \\ Y_2 &= S_1 S_2 \\ Y_n &= S_{n-2} S_{n-1} S_n; \quad n \geq 3. \end{aligned}$$

Thus, on the third step, the chain enters the closed irreducible subset of sequences of length 3. Setting $HHH \equiv 1$, $THT \equiv 2$, so $D = \{HHH, THT\} = \{1, 2\}$, we have from (2) that

$$\phi_{s1} = \frac{\mu_{s2} + \mu_{21} - \mu_{s1}}{\mu_{12} + \mu_{21}}.$$

Now we showed in Example 6.15 that $\mu_{s1} = 14$ and $\mu_{s2} = 10$. Also, in (b), we established that $\mu_{12} = 10$ and $\mu_{21} = 14$. Hence

$$\phi_{s1} = \frac{10 + 14 - 14}{10 + 14} = \frac{5}{12}$$

and $\phi_{s2} = \frac{7}{12}$ is the probability that THT occurs before HHH .

- (8) **Exercise** Show that for a fair coin the expected number of tosses to obtain HHH after HTH is 12, and the expected number required to obtain HTH after HHH is 8.
- (9) **Exercise** Show that the probability that HHH is observed before HTH is $\frac{3}{10}$.
- (10) **Exercise** Show that the probability that TTH is observed before HHH is $\frac{7}{10}$.
- (11) **Exercise** A fairground showman offers to play the following game. On payment of an entry fee of £1, a customer names a possible outcome of a sequence of 3 coin tosses; the showman then names another possible outcome, and a fair coin is tossed repeatedly until one of the named sequences is obtained in three successive throws. The player who named that sequence wins.
- Show that the probability that TTH beats HHH is $\frac{7}{8}$.
 - Show that the probability that TTH beats TTH is $\frac{2}{3}$.
 - Show that the probability that TTH beats THT is $\frac{7}{8}$.
 - Show that the probability that HTT beats TTH is $\frac{2}{3}$.
 - If the showman wants to make on average 30 p per game, what prize money should he offer:
 - if customers choose sequences at random?
 - if customers make the best possible choice?

Remark This game was named Penney–Ante by W. Penney in 1969.

9.17 Example: Poisson Processes

Let $X(t)$ be a Markov process taking values in the nonnegative integers; suppose that $X(t)$ is nondecreasing, with $X(0) = 0$. Suppose that as $h \rightarrow 0$ the transition rates satisfy

$$\begin{aligned} \mathbf{P}(X(t+h) = i+1 | X(t) = i) &= \lambda(t)h + o(h) \\ \mathbf{P}(X(t+h) = i | X(t) = i) &= 1 - \lambda(t)h + o(h), \end{aligned}$$

so that $X(t)$ changes its value by jumps of size one. Denote the times at which $X(t)$ jumps by T_1, T_2, T_3, \dots

- (a) Show that for fixed t , $X(t)$ has a Poisson mass function with parameter $\Lambda(t) = \int_0^t \lambda(u) du$. Hence, find the density of T_1 , the time of the first jump.
- (b) Find the joint density of T_1 and T_2 ; hence, find the conditional density of T_1 given T_2 .

Remark $X(t)$ is called a nonhomogeneous Poisson process with intensity (or rate) function $\lambda(t)$.

Solution (a) Let $p_n(t) = \mathbf{P}(X(t) = n)$. Then, by conditional probability,

$$p_n(t+h) = p_n(t)(1 - \lambda(t)h) + p_{n-1}(t)\lambda(t)h + o(h).$$

Hence, we obtain the forward equations in the usual manner as

$$(1) \quad \frac{dp_n(t)}{dt} = -\lambda(t)p_n(t) + \lambda(t)p_{n-1}(t); \quad n \geq 0,$$

where $p_{-1}(t) = 0$. Setting $G(z, t) = \sum_0^\infty z^n p_n(t)$, we find using (1) that

$$(2) \quad \frac{\partial G}{\partial t} = +\lambda(t)(z - 1)G.$$

Because $X(0) = 0$, we have $G(z, 0) = 1$, and so, by inspection, (2) has solution

$$(3) \quad G(z, t) = \exp\left((z - 1) \int_0^t \lambda(u) du\right).$$

This of course is the p.g.f. of the Poisson distribution with parameter $\Lambda(t)$, as required.

Now we note that

$$\mathbf{P}(T_1 > t) = \mathbf{P}(X(t) = 0) = G(0, t) = \exp(-\Lambda(t)).$$

(b) From (2) and (3), we can now see that for $w > t$,

$$\mathbf{E}(z^{X(w)} | X(t)) = z^{X(t)} \exp\left\{(z - 1) \int_t^w \lambda(u) du\right\},$$

and

$$\mathbf{E}\{y^{X(t)} z^{X(w)-X(t)}\} = \mathbf{E}(y^{X(t)})\mathbf{E}(z^{X(w)-X(t)}).$$

It follows that this nonhomogeneous Poisson process also has independent increments. Now

$$\begin{aligned} \mathbf{P}(T_1 > t, T_2 > w) &= \mathbf{P}(X(w) \in \{0, 1\}, X(t) = 0) \\ &= \mathbf{P}(X(t) = 0)\mathbf{P}(X(w) - X(t) \leq 1) \\ &\quad \text{by the independence of increments} \\ &= e^{-\Lambda(t)}[e^{-(\Lambda(w)-\Lambda(t))}(1 + \Lambda(w) - \Lambda(t))]. \end{aligned}$$

Hence, differentiating, T_1 and T_2 have joint density

$$f(t, w) = \lambda(t)\lambda(w)e^{-\Lambda(w)}; \quad 0 < t < w < \infty.$$

Integrating with respect to t shows that the density of T_2 is $f_{T_2}(w) = \lambda(w)\Lambda(w)e^{-\Lambda(w)}$, and so the conditional density is

$$f_{T_1|T_2}(t|w) = \frac{\lambda(t)}{\Lambda(w)}; \quad 0 < t < w.$$

- (4) **Exercise** Show that $\mathbf{P}(T_1 < \infty) = 1$ if and only if $\lim_{t \rightarrow \infty} \Lambda(t) = \infty$.
- (5) **Exercise** If $\lambda(t) = \lambda e^{-\lambda t}$ for $\lambda > 0$, show that $\lim_{t \rightarrow \infty} \mathbf{P}(X(t) = k) = 1/ek!$
- (6) **Exercise: Compound Poisson Process** Let $(Y_n; n \geq 1)$ be independent and identically distributed and independent of $X(t)$. Let $Z(t) = \sum_{n=1}^{X(t)} Y_n$. Show that $\mathbf{E}(e^{\theta Z(t)}) = \exp(\int_0^t \lambda(u) du (M(\theta) - 1))$, where $M(\theta) = \mathbf{E}(e^{\theta Y_1})$.
- (7) **Exercise: Doubly Stochastic Poisson Process** Suppose that $X(t)$ is a nonhomogeneous Poisson process with random intensity $\lambda(t)$; that is to say, for any realization of the process $X(t)$, $\lambda(t)$ is a realization of a random process $Y(t)$, where $\mathbf{E}(\exp[\theta \int_0^t Y(t) dt]) = M(\theta)$. Show that $X(t)$ has probability generating function $M(z - 1)$. Find the mean and variance of $X(t)$ in this case.

9.18 Example: Decay

Let $(T_n = \sum_{i=1}^n X_i; n \geq 1)$ be the partial sums of the independent exponential random variables $(X_i; i \geq 1)$ having parameter λ . A certain class of particles has the property that when freshly produced their (independent) lifetimes are exponential with parameter μ . At the ends of their lives, they disappear. At time T_n , a number Y_n of fresh particles is released into a chamber; the random variables $(Y_n; n \geq 0)$ are independent and identically distributed with p.g.f. $G_Y(z)$. At time t , the number of particles in the chamber is $N(t)$. Show that

$$(1) \quad \mathbf{E}(z^{N(t)}) = \exp\left(\lambda \int_0^t [G_Y((z-1)e^{-\mu v} + 1) - 1] dv\right).$$

Solution By construction, the batches arrive at the jump times of a Poisson process. Hence, given that k batches have arrived at time t , their arrival times are independently and uniformly distributed over $(0, t)$. For any particle in a batch of size Y that arrived at time U , the chance of survival to t is $e^{-\mu(t-U)}$ independently of all the others. Hence, given $U = u$, the p.g.f. of the number S of survivors of Y at t is

$$\mathbf{E}((ze^{-\mu(t-u)} + 1 - e^{-\mu(t-u)})^Y) = G_Y(ze^{-\mu(t-u)} + 1 - e^{-\mu(t-u)}).$$

Hence,

$$\mathbf{E}(z^S) = \mathbf{E}(\mathbf{E}(z^S|U)) = \frac{1}{t} \int_0^t G_Y((z-1)e^{-\mu v} + 1) dv.$$

Finally, recalling that the total number of particles at t are the survivors of a Poisson number of such batches we obtain (1).

- (2) **Exercise** What is $\mathbf{E}(z^{N(t)})$ when $G_Y(z) = z$? In this case, find $\lim_{t \rightarrow \infty} \mathbf{P}(N(t) = k)$.
- (3) **Exercise** In the case when $G_Y(z) = z$, show that $N(t)$ is a Markov process such that as $h \rightarrow 0$,

$$\begin{aligned} p_{i,i+1}(h) &= \lambda h + o(h) \\ p_{i,i-1}(h) &= i\mu h + o(h) \\ p_{ii}(h) &= 1 - \lambda h - i\mu h + o(h). \end{aligned}$$

Hence, obtain your answer to Exercise 2 by using the forward equations.

- (4) **Exercise** In the case when $G_Y(z) = z$, let t_n be the time when $N(t)$ makes its n th jump. Let $Z_n = N(t_n)$ be the imbedded Markov chain that records the successive different values of $N(t)$. Find the stationary distribution of Z_n .

9.19 Example: Disasters

A population evolves as follows. Immigrants arrive according to a Poisson process of rate ν . On arrival, each immigrant immediately starts a simple birth process with parameter λ independently of all other immigrants. Disasters occur independently of the population according to a Poisson process of rate δ ; when a disaster occurs, all individuals then in existence are annihilated. A disaster occurs at $t = 0$.

Let $X(t)$ denote the number of individuals in existence at time $t \geq 0$.

- (a) Show that $\lim_{t \rightarrow \infty} \mathbf{E}(X(t))$ is finite if and only if $\delta > \lambda$.
- (b) Find an expression for $\mathbf{E}(s^{X(t)})$.

Solution Because $X(t)$ is a Markov process, we could proceed by writing down forward equations. However, it is neater to use the properties of the Poisson process directly as follows. We start by assembling some facts established in earlier sections.

At time t , let $C(t)$ be the time that has elapsed since the most recent disaster. From Example 8.17, we recall that

$$(1) \quad \mathbf{P}(C(t) > x) = e^{-\delta x}; \quad 0 \leq x \leq t.$$

Now note that arrivals are a Poisson process independent of disasters, so given that $C(t) = x$, the number of subsequent arrivals up to time t is a Poisson random variable N , with parameter νx .

Next, we recall from Theorem 8.8.6 that conditional on $N = k$, these k arrivals are independently and uniformly distributed over the interval $(t - x, t)$, at times $t - Y_1, t - Y_2, \dots, t - Y_k$, say [where the Y_i are uniform on $(0, x)$].

Finally, we remember that given $Y_1 = y$, the expected number of descendants at time t from this arrival at $t - y$ is $e^{\lambda y}$; this is from (9.8.21). Now we remove the conditions one

by one. First, the expected number of descendants of an arrival at $t - Y_1$ is

$$(2) \quad \frac{1}{x} \int_0^x e^{\lambda y} dy = \frac{1}{\lambda x} (e^{\lambda x} - 1)$$

because Y_1 is uniform on $(0, x)$. Second, the expected number of descendants at t of the N arrivals during $(t - x, t)$ is, using (2),

$$(3) \quad \mathbf{E}(N) \frac{1}{\lambda x} (e^{\lambda x} - 1) = \frac{\nu}{\lambda} (e^{\lambda x} - 1).$$

Finally, using (1) and (3), we have

$$(4) \quad \mathbf{E}(X(t)) = \int_0^t \delta e^{-\delta x} \frac{\nu}{\lambda} (e^{\lambda x} - 1) dx + \frac{\nu}{\lambda} e^{-\delta t} (e^{\lambda t} - 1).$$

You can now see (if you want to) that, in more formal terms, what we have done is to say

$$\mathbf{E}(X) = \mathbf{E}(\mathbf{E}(\mathbf{E}[X|C, N, Y_1, \dots, Y_N]|C, N]|C)),$$

and then to successively evaluate the conditional expectations from the inside out.

So from (4), if $\lambda \geq \delta$, $\mathbf{E}(X(t)) \rightarrow \infty$ as $t \rightarrow \infty$, whereas if $\lambda < \delta$, $\mathbf{E}(X(t)) \rightarrow \frac{\nu}{\delta - \lambda}$. An expression for $\mathbf{E}(s^{X(t)})$ is found by following exactly the same sequence of successive conditional expectations. Thus, given that $C(t) = x$, $N(x) = k$, and $Y_1 = y_1$, this arrival initiates a simple birth process whose size at time y_1 has generating function

$$\frac{s e^{-\lambda y_1}}{1 - s + s e^{-\lambda y_1}} \quad \text{by Example 9.8.14.}$$

Hence, because Y_1 is uniformly distributed on $[0, x]$, the generating function of the number of descendants at time x of one arrival in $[0, x]$ is

$$(5) \quad \frac{1}{x} \int_0^x \frac{s e^{-\lambda y}}{1 - s + s e^{-\lambda y}} dy = -\frac{1}{\lambda x} \log(1 - s + s e^{-\lambda x}).$$

By independence, the generating function of the sum of k such independent arrivals is

$$(\log(1 - s + s e^{-\lambda x})^{-\frac{1}{\lambda x}})^k.$$

Next, we recall that $N(x)$ is Poisson with parameter νx , so that using conditional expectation again, the generating function of the descendants at t of the arrivals in $[t - x, t]$ is

$$(6) \quad \exp(\nu x (\log(1 - s + s e^{-\lambda x})^{-\frac{1}{\lambda x}} - 1)) = \frac{e^{-\nu x}}{(1 - s + s e^{-\lambda x})^{\frac{\nu}{\lambda}}}.$$

Now we recall from Example 8.17 that the current life (or age) of a Poisson process has density

$$f_{C(t)}(x) = \delta e^{-\delta x}; \quad 0 \leq x \leq t,$$

with $\mathbf{P}(C(t) = t) = e^{-\delta t}$. Hence, finally,

$$(7) \quad \mathbf{E}(s^{X(t)}) = \int_0^t \frac{\delta e^{-\delta x} e^{-\nu x}}{(1 - s + s e^{-\lambda x})^{\frac{\nu}{\lambda}}} dx + e^{-\delta t} \frac{e^{-\nu t}}{(1 - s + s e^{-\lambda t})^{\frac{\nu}{\lambda}}}.$$

- (8) **Exercise** Attempt to obtain (7) by writing down the forward equations for $\mathbf{P}(X(t) = n); n \geq 0$, and solving them.
- (9) **Exercise** Suppose that each immigrant gives rise to a simple birth and death process with parameters λ and μ . Show that $\lim_{t \rightarrow \infty} \mathbf{E}(X(t)) < \infty$ if and only if $\delta > \lambda - \mu$. [See 9.21].
- (10) **Exercise** Suppose that an ordinary immigration–death process with parameters ν and μ is subject to disasters. Show that the population size $X(t)$ has a stationary distribution with mean $\frac{\nu}{\delta + \mu}$. [Set $\lambda = 0$ in Exercise 9.]

9.20 Example: The General Birth Process

Let $(Y_n; n \geq 1)$ be a collection of independent exponentially distributed random variables such that Y_n has parameter λ_{n-1} . Let

$$T_n = \sum_{r=1}^n Y_r$$

and

$$N(t) = \max \{n: T_n \leq t\}.$$

The process $N(t)$ is a general birth process. Show that if

(1)
$$\sum_0^\infty \lambda_r^{-1} < \infty,$$

then for $t > 0$,

(2)
$$\mathbf{P}(N(t) < \infty) < 1.$$

Also, show that $\mathbf{E}(N(t) | N(t) < \infty)$ is finite or infinite, depending on whether $\sum_0^\infty r \lambda_r^{-1}$ converges or diverges.

Solution First, recall the often-used identity

(3)
$$\mathbf{P}(T_n \leq t) = \mathbf{P}(N(t) \geq n).$$

Let T_n have density $f_n(t)$ and moment generating function $M_n(\theta)$, and define

(4)
$$T = \lim_{n \rightarrow \infty} T_n = \sup \{t: N(t) < \infty\}.$$

Because the Y_n are independent and exponentially distributed, it follows that

(5)
$$M_n(\theta) = \prod_{r=0}^{n-1} (1 + \theta \lambda_r^{-1})^{-1}.$$

If (1) holds, then as $n \rightarrow \infty$ the infinite product converges (uniformly) to a nonzero limit $M(\theta)$. By the continuity theorem, this is the moment generating function of the density $f_T(t)$ of T . Hence, by (3),

$$\mathbf{P}(N(t) < \infty) = \mathbf{P}(T > t) = \int_t^\infty f_T(u) du < 1$$

for $t > 0$.

If (1) does not hold, then the product in (5) diverges to zero as $n \rightarrow \infty$, for $\theta \neq 0$, and $f_T(t) = 0$.

Furthermore, from (3),

$$(6) \quad p_n(t) = \mathbf{P}(N(t) = n) = \mathbf{P}(T_n \leq t) - \mathbf{P}(T_{n+1} \leq t) = \int_0^t f_n(u) du - \int_0^t f_{n+1}(u) du.$$

Now using (6), for $\theta < 0$, we have

$$(7) \quad \int_0^\infty e^{\theta t} p_n(t) dt = \frac{1}{\theta} (M_n(\theta) - M_{n+1}(\theta)) \quad \text{on integrating by parts} \\ = \frac{1}{\lambda_n} M_{n+1}(\theta) \quad \text{using (5).}$$

Because $M_n(\theta)$ converges uniformly to $M(\theta)$, we can use the inversion theorem on each side of (7) to find that, as $n \rightarrow \infty$,

$$(8) \quad \lambda_n p_n(t) \rightarrow f_T(t).$$

Now

$$\mathbf{E}(N(t) | N(t) < \infty) = \frac{\sum_{n=0}^{\infty} n p_n(t)}{\sum_{n=0}^{\infty} p_n(t)},$$

which converges or diverges with $\sum n p_n(t)$. Using (8), it follows that $\mathbf{E}(N(t) | N(t) < \infty) < \infty$ if and only if $\sum_{n=0}^{\infty} n \lambda_n^{-1} < \infty$.

(9) **Exercise** Write down the forward equations for $p_n(t)$ and deduce (7) directly from these.

(10) **Exercise** Deduce from (7) that $p_n(t) = \frac{1}{\lambda_n} \sum_{i=0}^n a_i \lambda_i e^{-\lambda_i t}$, where

$$a_i = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{\lambda_j}{\lambda_j - \lambda_i}.$$

(11) **Exercise** Show that if $\lambda_n = n(\log n)^\gamma$; $\gamma > 1$ then for any $\beta > 0$, $\mathbf{E}([N(t)]^\beta | N(t) < \infty) = \infty$.

9.21 Example: The Birth–Death Process

Let the Markov process $X(t)$ represent the number of individuals in a population at time t . During any interval $(t, t + h)$, any individual alive at t may die with probability $\mu h + o(h)$, or split into two individuals with probability $\lambda h + o(h)$. All individuals act independently in these activities. Write down the forward equations for $p_n(t) = \mathbf{P}(X(t) = n)$ and show that if $X(0) = I$, then

$$(1) \quad \mathbf{E}(s^{X(t)}) = \begin{cases} \left(\frac{s + \lambda t(1-s)}{1 + \lambda t(1-s)} \right)^I & \text{if } \lambda = \mu \\ \left(\frac{\mu \exp(t(\lambda - \mu)) + \theta(s)}{\lambda \exp(t(\lambda - \mu)) + \theta(s)} \right)^I & \text{if } \lambda \neq \mu, \end{cases}$$

where $\theta(s) = (\lambda s - \mu)/(1 - s)$.

Solution Because individuals act independently, the probability of no change during $(t, t + h)$ when $X(t) = k$ is

$$p_{kk}(h) = (1 - \lambda h - \mu h + o(h))^k = 1 - (\mu + \lambda)kh + o(h).$$

Similarly, the probability of just one split and no deaths among k individuals during $(t, t + h)$ is

$$p_{k,k+1}(h) = k\lambda h(1 - \lambda h - \mu h + o(h))^{k-1} = k\lambda h + o(h),$$

and likewise the chance of just one death is $p_{k,k-1}(h) = k\mu h + o(h)$. Other transitions have probabilities that are all $o(h)$ as $h \rightarrow 0$, and so by conditional probability

$$p_k(t + h) = h\lambda(k - 1)p_{k-1}(t) + h\mu(k + 1)p_{k+1}(t) + (1 - (\lambda + \mu)kh)p_k(t) + o(h).$$

The forward equations now follow as usual, giving

$$\frac{dp_k(t)}{dt} = \lambda(k - 1)p_{k-1}(t) + \mu(k + 1)p_{k+1}(t) - (\lambda + \mu)kp_k(t),$$

with the convention that $p_{-1}(t) = 0$. Defining $G(s, t) = \mathbf{E}(s^{X(t)})$ and differentiating G with respect to s , shows that

$$(2) \quad \frac{\partial G}{\partial t} = \lambda s^2 \frac{\partial G}{\partial s} + \mu \frac{\partial G}{\partial s} - (\lambda + \mu)s \frac{\partial G}{\partial s} = (\lambda s - \mu)(s - 1) \frac{\partial G}{\partial s}.$$

Because $X(0) = I$, we have $G(s, 0) = s^I$, and it is straightforward but dull to verify that (1) satisfies (2) and the initial condition $X(0) = I$.

(3) **Exercise** Let η be the probability that the population ever falls to zero. Show that

$$\eta = \begin{cases} 1 & \text{if } \mu \geq \lambda \\ \left(\frac{\mu}{\lambda}\right)^I & \text{if } \lambda > \mu. \end{cases}$$

(4) **Exercise** Let T be the time until $X(t)$ first takes the value zero. Show that if $X(0) = 1$,

$$\mathbf{E}(T|T < \infty) = \begin{cases} \frac{1}{\lambda} \log \frac{\mu}{\mu - \lambda}; & \lambda < \mu \\ \frac{1}{\mu} \log \frac{\lambda}{\lambda - \mu}; & \lambda > \mu. \end{cases}$$

(5) **Exercise** Let $X(0) = 1$ and define $z(t) = \mathbf{P}(X(t) = 0)$. Show that $z(t)$ satisfies

$$\frac{dz}{dt} = \mu - (\lambda + \mu)z(t) + \lambda(z(t))^2.$$

Hence, find $z(t)$. What is $\mathbf{P}(X(t) = 0|X(s) = 0)$ for $0 < t < s$?

(6) **Exercise** Suppose that $X(0) = 1$ and $\lambda < \mu$. Show that

$$\lim_{t \rightarrow \infty} \mathbf{P}(X(t) = k|X(t) > 0) = \left(\frac{\lambda}{\mu}\right)^k \left(1 - \frac{\lambda}{\mu}\right).$$

- (7) **Exercise** Suppose that new individuals join the population at the instants of a Poisson process with parameter ν (independently of the birth and death process). Write down the forward equations for the process. Deduce that if $\lambda < \mu$ the stationary distribution is

$$\pi_k = \frac{1}{k!} \left(1 - \frac{\lambda}{\mu}\right)^{\frac{\nu}{\lambda}} \left(\frac{\lambda}{\mu}\right)^k \left(\frac{\nu}{\lambda} + k - 1\right) \left(\frac{\nu}{\lambda} + k - 2\right) \dots \frac{\nu}{\lambda}.$$

What is the mean of this distribution?

9.22 Example: Wiener Process with Drift

Let $D(t)$ be the drifting standard Wiener process, $D(t) = \mu t + W(t)$.

- (a) Show that $M(t)$ is a martingale, where

$$M(t) = \exp\{\lambda D(t) - \frac{1}{2}\lambda^2 t - \lambda\mu t\}.$$

- (b) Let T_b be the first passage time of $D(t)$ to $b > 0$. Show that

$$\mathbf{E}e^{-\theta T_b} = \exp\{b(\mu - \sqrt{\mu^2 + 2\theta})\}.$$

Solution (a) By definition $D(t) = \mu t + W(t)$, so $M(t) = \exp\{\lambda W(t) - \frac{1}{2}\lambda^2 t\}$, which we know to be a martingale from 9.9.27.

(b) Because T_b is a stopping time, $1 = \mathbf{E}M(T_b \wedge t)$. But if λ is so large that $\lambda\mu + \frac{1}{2}\lambda^2 > 0$, then $0 \leq M(T_b \wedge t) \leq e^{\lambda b}$. Let $t \rightarrow \infty$ and use the dominated convergence theorem 5.9.7 to give

$$1 = \mathbf{E}M(T_b) = e^{\lambda b} \mathbf{E}e^{-T_b(\frac{1}{2}\lambda^2 + \lambda\mu)}.$$

Now setting $\theta = \lambda\mu + \frac{1}{2}\lambda^2$ and choosing the larger root of the quadratic (which yields a moment generating function) gives the result.

- (1) **Exercise** Find $\mathbf{P}(T_b < \infty)$ in the two cases $\mu > 0$ and $\mu < 0$.
 (2) **Exercise** Show that $e^{-2\mu D(t)}$ is a martingale.
 (3) **Exercise** Let T be the time at which $D(t)$ first hits a or b , where $a < 0$ and $b > 0$. Show that the probability that $D(t)$ hits b first is

$$\mathbf{P}(D(T) = b) = \frac{1 - e^{-2a\mu}}{e^{-2b\mu} - e^{-2a\mu}}.$$

- (4) **Exercise** Show that if $\mu < 0$, then

$$\mathbf{P}(\max_{t \geq 0} D(t) \geq b) = e^{2\mu b}.$$

(Note: In the remaining Exercises (5)–(9), we consider the standard Wiener process with no drift in which $\mu = 0$.)

- (5) **Exercise** Let $W(t)$ be the Wiener process and T_b the first passage time of $W(t)$ to $b > 0$. Show that $\mathbf{E}e^{-\theta T_b} = \exp(-\sqrt{2\theta}b)$.

Remark We can deduce from (5) that $\mathbf{P}(T_b < \infty) = 1$, but see also the next Exercise (6).

- (6) **Exercise** Use Example 9.9.33 to show that $\mathbf{P}(T_b < \infty) = 1$.
- (7) **Exercise** Let $X(t)$ and $Y(t)$ be independent Wiener processes, and let T_b be the first passage time of $X(t)$ to $b > 0$. Use conditional expectation and (5) to show that $\mathbf{E}e^{i\theta Y(T_b)} = e^{-|\theta|b}$, and deduce that $Y(T_b)$ has a Cauchy density.
(Hint for the final part: Look at Example 8.24.)
- (8) **Exercise** Use the fact that the density of T_b is given in Corollary 9.9.25 to calculate the density of $Y(T_b)$ directly.
- (9) **Exercise** Let $c > b$. Explain why $Y(T_b)$ is independent of $Y(T_c) - Y(T_b)$. Now recall Example 9.9.11 and use it to deduce that $Y(T_b)$ has the same density as $bY(T_1)$. Finally, use this and the fact that $Y(T_b)$ has the same density as $-Y(T_b)$ to conclude that $\mathbf{E}e^{i\theta Y(T_b)} = e^{-K|\theta|}$, for some constant K .

9.23 Example: Markov Chain Martingales

Let $(X_n; n \geq 0)$ be a Markov chain with transition probabilities p_{ij} , and suppose that the function $v(\cdot, \cdot)$ is such that

$$\sum_j p_{ij}v(j, n + 1) = \lambda v(i, n), \quad \lambda \neq 0.$$

Show that $\lambda^{-n}v(X_n, n)$ is a martingale with respect to X_n , provided that $\mathbf{E}|v(X_n, n)| < \infty$.

Solution Using the Markov property of X_n ,

$$\begin{aligned} \mathbf{E}(\lambda^{-(n+1)}v(X_{n+1}, n + 1)|X_0, \dots, X_n) &= \lambda^{-(n+1)} \sum_j p_{X_n j}v(j, n + 1) \\ &= \lambda^{-(n+1)}\lambda v(X_n, n). \end{aligned}$$

The result follows, on noting that $\mathbf{E}|v(X_n, n)| < \infty$.

- (1) **Exercise** Let X_n be a Markov chain with state space $\{0, 1, \dots, b\}$, such that X_n is also a martingale. Show that 0 and b are absorbing states, and that if absorption occurs with probability one in finite time, then the probability of absorption at b is X_0/b .
- (2) **Exercise** Let X_n be a Markov chain with state space $\{0, 1, \dots, b\}$ and transition probabilities p_{ij} ; and suppose that the bounded function $v(\cdot)$ is such that $v(0) = 0$, $v(b) = 1$, and

$$v(i) = \sum_{j \in S} p_{ij}v(j), \quad i \in S.$$

If 0 and b are absorbing states, show that if absorption occurs in finite time with probability one, then the probability of absorption at b is $v(X_0)$.

- (3) **Exercise** Let X_n be a birth–death process on the nonnegative integers, with transition probabilities

$$\begin{aligned} p_{i, i+1} &= p_i, & i > 0 \\ p_{i, i-1} &= q_i, & i > 0 \\ p_{ii} &= r_i, & i \geq 0, \quad \text{where } r_0 = 1. \end{aligned}$$

Define

$$v(x) = \sum_{r=1}^x \prod_{i=1}^{r-1} \frac{q_i}{p_i}, \quad x \geq 2,$$

while $v(0) = 0$ and $v(1) = 1$. Show that $v(X_n)$ is a martingale. Deduce that the probability of hitting b before a , given $X_0 = x < b$, is $\frac{v(x)-v(a)}{v(b)-v(a)}$. Deduce that the process is persistent if and only if $\sum_{i=1}^{\infty} \prod_{j=1}^{i-1} \frac{q_j}{p_j}$ diverges.

9.24 Example: Wiener Process Exiting a Strip

Let $W(t)$ be the standard Wiener process and let T be the first time at which $W(t)$ hits a or b , where $a < 0$ and $b > 0$. Show that $\mathbf{E}T = -ab$.

Solution We know from Theorem 9.9.27 that $W(t)^2 - t$ is a martingale. Hence, for any finite integer n , because T is a stopping time,

$$(1) \quad 0 = \mathbf{E}W(T \wedge n)^2 - \mathbf{E}(T \wedge n).$$

Because $W(t)^2 \leq a^2 + b^2$, we can let $n \rightarrow \infty$ in (1) to obtain, by using Example 9.9.33,

$$\begin{aligned} \mathbf{E}T &= \lim_{n \rightarrow \infty} \mathbf{E}(T \wedge n) = \lim_{n \rightarrow \infty} \mathbf{E}(W(T \wedge n)^2) \\ &= \mathbf{E}W(T)^2 \\ &= a^2 \mathbf{P}(W(T) = a) + b^2 \mathbf{P}(W(T) = b) \\ &= \frac{a^2 b}{b - a} - \frac{ab^2}{b - a} \\ &= -ab. \end{aligned}$$

(2) **Exercise** Let T_b be the first passage time of $W(t)$ to $b \neq 0$. Show that $\mathbf{E}T_b = \infty$.

(3) **Exercise** Use the martingales 9.9.29–9.9.31 to show that

$$3\mathbf{E}T^2 = 3a^2b^2 - ab(a^2 + b^2) \quad \text{and} \quad 3\text{var}T = -ab(a^2 + b^2).$$

(4) **Exercise** Use the martingale $e^{\theta W(t) - \frac{1}{2}\theta^2 t} = M_\theta$ to show that, when $a = -b$,

$$\mathbf{E}e^{-sT} = [\cosh(a\sqrt{2s})]^{-1}.$$

(Hint: Show that $M_\theta + M_{-\theta}$ is a martingale.)

(5) **Exercise** Use the result of Example 9.23 on first passage times to show that, for any $a < 0$, $b > 0$,

$$\mathbf{E}e^{-\theta T} = \frac{\sinh(\sqrt{2\theta}b) - \sinh(\sqrt{2\theta}a)}{\sinh[\sqrt{2\theta}(b-a)]}$$

(Hint: Write

$$\mathbf{E}e^{-\theta T_b} = \mathbf{E}[e^{-\theta T_b} I\{T_a < T_b\}] + \mathbf{E}[e^{-\theta T_b} I\{T_b < T_a\}]$$

and use the Markov property at T_a . Then, do likewise for $\mathbf{E}e^{-\theta T_a}$.)

9.25 Example: Arcsine Law for Zeros

Show that the probability that the Wiener process has no zero in (s, t) is

$$\frac{2}{\pi} \sin^{-1} \sqrt{\frac{s}{t}} = \frac{2}{\pi} \arcsin \sqrt{\frac{s}{t}}.$$

Solution Let Z be the event that $W(t)$ does have at least one zero in (s, t) . Recall that T_w is the first passage time of $W(t)$ to w , with density f_T given in Corollary 9.9.25. By the symmetry of the Wiener process, if $W(s) = w$, then

$$\mathbf{P}(Z) = \mathbf{P}(T_w \leq t - s) = \mathbf{P}(T_{-w} \leq t - s).$$

Therefore, conditioning on $W(s)$, we have

$$\begin{aligned} \mathbf{P}(Z) &= 2 \int_{w=0}^{\infty} \int_{u=0}^{t-s} f_T(u) f_{W(s)}(-w) \, du \, dw \\ &= \frac{1}{\pi \sqrt{s}} \int_{u=0}^{t-s} u^{-3/2} \int_{w=0}^{\infty} w \exp \left\{ -\frac{1}{2} w^2 \frac{u+s}{us} \right\} \, dw \, du \\ &= \frac{\sqrt{s}}{\pi} \int_0^{t-s} \frac{du}{(u+s)\sqrt{u}} \\ &= \frac{2}{\pi} \tan^{-1} \sqrt{\frac{t}{s} - 1}, \quad \text{on setting } u = sv^2 \\ &= \frac{2}{\pi} \cos^{-1} \sqrt{\frac{s}{t}}, \end{aligned}$$

using the right-angled triangle with sides $\sqrt{\frac{s}{t}}$, $\sqrt{1 - \frac{s}{t}}$, 1.

Finally, the required probability is

$$\mathbf{P}(Z^c) = 1 - \mathbf{P}(Z) = \frac{2}{\pi} \sin^{-1} \sqrt{\frac{s}{t}},$$

on using the same right-angled triangle with sides $\sqrt{\frac{s}{t}}$, $\sqrt{1 - \frac{s}{t}}$, and 1.

(1) **Exercise** Let V_1 be the time of the last zero of $W(t)$ before t , and V_2 the time of the first zero after t . Show that

(i) $\mathbf{P}(V_2 \leq s) = \frac{2}{\pi} \cos^{-1} \sqrt{\frac{t}{s}}, t < s.$

(ii) $\mathbf{P}(V_1 < s, V_2 > v) = \frac{2}{\pi} \sin^{-1} \sqrt{\frac{s}{v}}, s < t < v.$

(2) **Exercise** Show that the probability that the Brownian bridge has no zeros in $(s, t), 0 < s < t < 1$, is

$$\frac{2}{\pi} \cos^{-1} [(t-s)/[t(1-s)]]^{1/2}.$$

(3) **Exercise** If $M(t) = \sup_{0 \leq s \leq t} W(s)$, argue that

$$\mathbf{P}(M(t) \leq y | T_c = s) = \mathbf{P}(M(t-s) \leq y - c), \quad s \leq t.$$

Deduce that $M(t)$ and T_c have the joint density

$$f_{M,T}(y,u) = \frac{c}{\pi u \sqrt{u(t-u)}} \exp \left\{ -\frac{1}{2} \frac{(y-c)^2}{t-u} - \frac{1}{2} \frac{c^2}{u} \right\},$$

- (4) **Exercise** Let $U(t)$ be the time at which $W(t)$ attains its maximum in $[0, t]$. [It can be shown that $U(t)$ exists and is unique with probability 1.] Use the previous exercise, and the fact that $U(t) = T_x$ on the event $M(t) = x$ to show that $M(t)$ and $U(t)$ have joint density

$$f_{M,U}(x,u) = \frac{x}{\pi u \sqrt{u(t-u)}} \exp \left(-\frac{x^2}{2u} \right).$$

Deduce that $U(t)$ satisfies

$$\mathbf{P}(U \leq u) = \frac{2}{\pi} \sin^{-1} \sqrt{\frac{u}{t}}, \quad 0 \leq u \leq t.$$

9.26 Example: Option Pricing: Black–Scholes Formula

In Example 9.9.36, we gave the fair price at $t = 0$ for a European call option with exercise time T as

$$v = \mathbf{E}\{e^{-rT}(S(T) - K)^+\},$$

where the stock price $S(t)$ is assumed to be a geometric Wiener process of the form

$$(1) \quad S(t) = S(0)\{\mu t + \sigma W(t)\}$$

and $\mu + \frac{1}{2}\sigma^2 = r$. Show that v can be written explicitly as

$$(2) \quad v = S(0)\Phi(H) - Ke^{-rT}\Phi(H - \sigma\sqrt{T}),$$

where $\Phi(x)$ is the standard normal distribution function and

$$H = \{(r + \frac{1}{2}\sigma^2)T + \log[S(0)/K]\}/\{\sigma\sqrt{T}\}.$$

Solution Consider a random variable Z with the normal $N(\gamma, \tau^2)$ density. We have

$$\begin{aligned} \mathbf{E}(ae^Z - K)^+ &= \int_{\log K/a}^{\infty} (ae^z - K) \frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{-(z-\gamma)^2}{2\tau^2}\right) dz \\ &= \int_{\alpha}^{\infty} (ae^{\gamma+\tau y} - K) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) dy, \\ &\quad \text{where } y = \frac{z-\gamma}{\tau}, \text{ and } \alpha = \frac{\log K/a - \gamma}{\tau}, \\ &= ae^{\gamma+\frac{1}{2}\tau^2} \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\tau)^2\right) dy - K\Phi(-\alpha) \\ &= ae^{\gamma+\frac{1}{2}\tau^2} \Phi(\tau - \alpha) - K\Phi(-\alpha). \end{aligned}$$

For the problem in question, we can write $S(T) = ae^Z$, where $a = S(0)$ and Z is normal $N((r - \frac{1}{2}\sigma^2)T, \sigma^2T)$. Inserting these values of a , γ , and τ in the above, shows that

$$\begin{aligned} v &= \mathbf{E}e^{-rT}(S(T) - K)^+ = e^{-rT}\{S(0)e^{(r-\frac{1}{2}\sigma^2)T+\frac{1}{2}\sigma^2T}\Phi(\tau - \alpha) - K\Phi(-\alpha)\} \\ &= S(0)\Phi\left(\frac{\tau^2 - \log(\frac{K}{a}) + \gamma}{\tau}\right) - Ke^{-rT}\Phi(-\alpha) \\ &= S(0)\Phi\left(\frac{\sigma^2T + rT - \frac{1}{2}\sigma^2T + \log\frac{S(0)}{K}}{\sigma\sqrt{T}}\right) - Ke^{-rT}\Phi(H - \sigma\sqrt{T}), \end{aligned}$$

as required.

- (3) **Exercise** Show that the value v of the option given in (2) is an increasing function of each of $S(0)$, T , r and σ , but is a decreasing function of K .
- (4) **Exercise** The ‘‘American Call option’’ differs from the European call in one respect: It may be exercised by the buyer at any time up to the expiry time T . Show that the value of the American call is the same as that of the corresponding European call and that there is no advantage to the holder in exercising it prior to the expiry time T .

PROBLEMS

- 1 Let $(X_n; n \geq 1)$ be a collection of independent identically distributed nonnegative random variables. Define:
 - (i) $S_n = \sum_{i=1}^n X_i$.
 - (ii) $M_n = \max\{X_1, X_2, \dots, X_n\}$.
 - (iii) $L_n = \min\{X_1, X_2, \dots, X_n\}$.
 - (iv) $K_n = X_n + X_{n-1}$.
 - (a) Which of the sequences X , S , M , L , K are Markov chains?
 - (b) For those that are, find the transition probabilities.
- 2 Classify the chains in Problem 1; that is to say, show whether the states are persistent, null, periodic.
- 3 Can a reversible chain be periodic?
- 4 Let $(X_n; n \geq 1)$ and $(Y_n; n \geq 1)$ be independent irreducible Markov chains, and set $Z_n = (X_n, Y_n); n \geq 1$.
 - (a) Is Z_n irreducible?
 - (b) If X and Y are reversible and also aperiodic, show that Z is reversible.
- 5 Let X be a Markov chain. Show that the sequence $(X_i; i \geq 0)$ conditional on $X_m = r$ still has the Markov property.
- 6 Show that Definition 9.1.1 is equivalent to each of (9.1.6), (9.1.7), and (9.1.8) as asserted.
- 7 Let Y_n be the number of heads shown in n tosses of a coin. Let $Z_n = Y_n$ modulo 10. Show that $(Z_n; n \geq 0)$ is a Markov chain; find its transition probabilities and stationary distribution.
- 8 Let $(S_n; n \geq 0)$ be a simple random walk with $S_0 = 0$; show that $Y_n = |S_n|$ is a Markov chain.
- 9 Let $(X_n; n \geq 1)$ be a Markov chain. Show that if $g(X_n)$ is any function of X_n , then $\mathbf{E}(\mathbf{E}(g(X_{n+m})|X_n)|X_r) = \mathbf{E}(g(X_{n+m})|X_r)$, for $r \leq n$.
- 10 Let $(u_n; n \geq 0)$ be a sequence defined by $u_0 = 1$ and $u_n = \sum_{k=1}^n f_k u_{n-k}$, where $f_k > 0$ and $\sum_{k=1}^\infty f_k < 1$.
 - (a) Show that v_n defined by $v_n = \rho^n u_n; n \geq 0$, is a renewal sequence as defined in Example 9.13, if $\sum_{n=1}^\infty \rho^n f_n = 1$.
 - (b) Show that as $n \rightarrow \infty$, for some constant c , $\rho^n u_n \rightarrow c$.

- 11** **Murphy's Law** Let $(X_n; n \geq 1)$ be an irreducible aperiodic persistent chain. Let $s = (s_1, \dots, s_m)$ be any finite sequence of states of the chain such that $p_{s_1 s_2} p_{s_2 s_3} \dots p_{s_{m-1} s_m} > 0$. Show that with probability 1 the sequence s occurs in finite time. Explain the implications.
- 12** Let $(X_n; n \geq 0)$ be a Markov chain. Show that for any constant d the sequence $(X_{nd}; n \geq 0)$ is a Markov chain.
- 13** Let A be a subset of the states of a regular chain X . Let $T_1 < T_2 < T_3 < \dots$ be the successive times at which the chain visits A . Show that $(X_{T_r}; r \geq 1)$ is a Markov chain.
- 14** Let $(X_n; n \geq 0)$ and $(Y_n; n \geq 0)$ be Markov chains with the same state space S , and distinct transition matrices p_{ij}^X and p_{ij}^Y . Let $(W_n; n \geq 0)$ be a process defined on S with transition probabilities

$$q_{ij}(n) = \frac{1}{2} \left(p_{ij}^X(n) + p_{ij}^Y(n) \right).$$

Show that $q_{ij} \geq 0$ and $\sum_j q_{ij} = 1$, but that $(W_n; n \geq 0)$ is not a Markov chain in general.

- 15** **Truncation** Let $(X_n; n \geq 0)$ be an irreducible Markov chain with state space S , transition probabilities p_{ij} , and stationary distribution $(\pi_i; i \in S)$. Let A be some subset of S , and suppose that a new chain Y is formed by banning transitions out of A . That is to say, Y has transition probabilities q_{ij} , where for $i \in A$, $q_{ij} = p_{ij}$ for $j \in A$ and $j \neq i$, and $q_{ii} = p_{ii} + \sum_{j \in A^c} p_{ij}$. Show that if X is reversible in equilibrium, then so is Y , and write down the stationary distribution of Y .
- 16** Let X_n and Y_n be independent simple random walks. Let Z_n be (X_n, Y_n) truncated as in Problem 15 to the region $x \geq 0, y \geq 0, x + y \leq a$. Find the stationary distribution of Z_n .
- 17** Let $(X_n; n \geq 0)$ be a Markov chain with state space S . For each $n \geq 0$ independently, X_n is replaced by $s \in S$ with probability p . Is the new sequence a Markov chain?
- 18** At each time $n = 0, 1, 2, \dots$ a number Y_n of particles is injected into a chamber, where $(Y_n; n \geq 0)$ are independent Poisson random variables with parameter λ . The lifetimes of particles are independent and geometric with parameter p . Let X_n be the number of particles in the chamber at time n . Show that X_n is a Markov chain; find its transition probabilities and the stationary distribution.
- 19** Let $(f_k; k \geq 0)$ be a probability mass function. Let the irreducible Markov chain X have transition probabilities,

$$p_{jk} = f_{k-j+1} \quad \text{if } k - j + 1 \geq 0, j \geq 1$$

and $p_{0k} = p_{1k}$. Show that X is recurrent and nonnull if $\sum_{k=1}^{\infty} k f_k < 1$.

- 20** Let $(f_k; k \geq 0)$ be a probability mass function. Suppose the Markov chain X has transition probabilities

$$p_{jk} = \begin{cases} f_{j-k+1} & \text{for } k > 0, j - k + 1 \geq 0 \\ \sum_{i=j+1}^{\infty} f_i & \text{for } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Show that X is recurrent and nonnull if $\sum_{k=1}^{\infty} k f_k > 1$.

- 21** **Lumping** Let X have state space S and suppose that $S = \cup_k A_k$, where $A_i \cap A_j = \emptyset$ for $i \neq j$. Let $(Y_n; n \geq 0)$ be a process that takes the value y_k whenever the chain X lies in A_k . Show that Y is also a Markov chain if $p_{i_1 j} = p_{i_2 j}$ for any i_1 and i_2 in the same set A_k .
- 22** **Markov Times** Let X be a Markov chain. Let T be a positive random variable such that $\mathbf{P}(T = t | X_0, \dots, X_t)$ is either zero or one. T is called a *Markov time*. Show that the Markov property is preserved at T .
- 23** Let S_n be the random walk, such that $\mathbf{P}(S_{n+1} - S_n = 2) = p$, and $\mathbf{P}(S_{n+1} - S_n = -1) = q$, where $p + q = 1$.

If the origin is a retaining barrier, show that equilibrium is possible with $S_n \geq 0$ if $p < \frac{1}{3}$ and that, in this case, the stationary distribution has p.g.f.,

$$\pi(s) = \frac{(1 - 3p)(s - 1)}{s - q - ps^3}.$$

24 Let $X(t)$ be the two-state chain in continuous time, $t \in \mathbb{R}$, $X(t) \in \{0, 1\}$, having stationary distribution $\{\pi_0, \pi_1\}$.

(a) Show that as $\tau \rightarrow \infty$

$$\mathbf{P}(X(0) = 1 | X(-\tau), X(\tau)) \rightarrow \pi_1.$$

(b) Find $\text{cov}(X(s), X(s + t))$; $t > 0$.

(c) What is $\lim_{s \rightarrow \infty} \text{cov}(X(s), X(s + t))$?

25 Let $N(t)$ and $M(t)$ be independent Poisson processes with parameters λ and μ , respectively.

(a) Is $N(t) + M(t)$ a Poisson process?

(b) Is either of $\min\{N(t), M(t)\}$ or $\max\{N(t), M(t)\}$ a Poisson process?

26 Let $N(t)$ be a nonhomogeneous Poisson process with rate $\lambda(t)$. Find $\text{cov}(N(s), N(s + t))$; $t > 0$.

27 Mosquitoes land on your neck at the jump times of a Poisson process with parameter $\lambda(t)$ and each bites you with probability p independently of the decisions of the others. Show that bites form a Poisson process with parameter $p\lambda(t)$.

28 Let $X(t)$ be a Markov chain with transition probabilities $p_{ij}(t)$ and stationary distribution π . Let $(T_n; n \geq 0)$ be the jump times of a Poisson process independent of $X(t)$. Show that the sequence $Y_n = X(T_n)$ is a Markov chain with the same stationary distribution as $X(t)$.

29 Find the mean and variance of the size $X(t)$ of the population in the birth–death process of Example 9.21.

30 **A Nonhomogeneous Chain** Let $X(t)$ be a Markov chain with $X(0) = I$ and such that, as $h \rightarrow 0$,

$$\mathbf{P}(X(t + h) = k + 1 | X(t) = k) = \frac{1 + \mu k}{1 + \mu t} h + o(h)$$

and

$$\mathbf{P}(X(t + h) = k | X(t) = k) = 1 - \frac{1 + \mu k}{1 + \mu t} h + o(h).$$

Show that $G = \mathbf{E}(s^{X(t)})$ satisfies

$$\frac{\partial G}{\partial t} = \frac{s - 1}{1 + \mu t} \left(G + \mu s \frac{\partial G}{\partial s} \right).$$

Hence, find $\mathbf{E}(X)$ and $\text{var}(X)$.

31 **Truncation Again** Let $(X_n; n \geq 0)$ be an irreducible Markov chain, with state space S , stationary distribution $(\pi_i; i \in S)$, and transition probabilities p_{ij} . Let A be some subset of S , and suppose that $(Z_n; n \geq 0)$ is a Markov chain with state space A and transition probabilities

$$q_{ij} = \frac{p_{ij}}{p_{iA}} \quad \text{for } i, j \in A, \quad \text{where } p_{iA} = \sum_{j \in A} p_{ij}.$$

If X is reversible, show that Z is reversible with stationary distribution given by

$$v_i = \pi_i p_{iA} / \sum_{i \in A} \pi_i p_{iA}.$$

- 32** “Motto” is a coin-tossing game at the start of which each player chooses a sequence of three letters, each of which is either H or T (his “motto”). A fair coin is then tossed repeatedly, and the results recorded as a sequence of H s and T s (H for “heads,” T for “tails”). The winner is the first player whose motto occurs as three consecutive letters in this sequence. Four players A, B, C, D choose as their mottoes, respectively, $HHT, THH, TTH,$ and HTT . Show that if only A and B take part in a game then B has probability $\frac{3}{4}$ of winning. With what probability does C win if he plays a game with B as the only opponent? If all four players take part simultaneously, what are the respective probabilities of each player winning?
(You may assume that if a fair coin is tossed repeatedly then with probability 1 any motto will occur eventually.)
- 33** Let $N(t)$ be a nonhomogeneous Poisson process. Show that, conditional on $N(t) = k$, the times T_1, \dots, T_k of the events have conditional joint density

$$k! \prod_{i=1}^k \left\{ \frac{\lambda(t_i)}{\Lambda(t)} \right\}, 0 \leq t_1 \leq \dots \leq t_k \leq t.$$

- 34** Show that (a) the Brownian bridge, (b) the reflected Wiener process $|W(t)|$, and (c) the Ornstein–Uhlenbeck process all satisfy the Markov property.
Explain why the integrated Wiener process $R(t) = \int_0^t W(u)du$ does not have this property.
- 35** Show that the transition density $f(s, x; t, y)$ of the Wiener process satisfies the Chapman–Kolmogorov equations

$$f(s, x; u, z) = \int_{\mathbb{R}} f(s, x; t, y) f(t, y; u, z) dy,$$

$$s < t < u.$$

- 36** Let $W(t)$ be the Wiener process. Show that for $s < t < u$,

$$\mathbf{E}(W(t)|W(u), W(s)) = [(t - s)W(u) + (u - t)W(s)]/(u - s)$$

and

$$\text{var}(W(t)|W(u), W(s)) = [u(t - s) + t(s - t)]/(u - s).$$

(Hint: Use the result of Exercise 4 in Example 8.20)

Deduce that the conditional correlation is, for $u = 1$,

$$\rho(W(s), W(t)|W(1)) = \left\{ \frac{s(1 - t)}{t(1 - s)} \right\}^{1/2}.$$

- 37** The random walk $X(n)$ takes place on the non-negative integers. From any nonzero value r , it steps to one of $\{0, 1, 2, \dots, r + 1\}$ with probability $1/(r + 2)$. From 0 it surely steps to 1.
- (a) Find the stationary distribution, and deduce that the expected number of steps to reach 0 starting from 1 is $2(e - 1)$.
- (b) Show that in equilibrium, if the walk is at 0, the probability that it arrived there from r is $2/\{(r + 2)r!\}$.
- (c) If the walk starts at 1, find the probability that it visits 0 before visiting $r + 1$.
- 38** I walk to and from work, and I have a total of m umbrellas at home or in my office. If it is raining when I set off either way, I take an umbrella if I have one to take. For any journey, it is raining independently with probability $p = 1 - q$. Let $U(n)$ be the number of umbrellas available to hand when I start the n th journey; ignore the chance of rain starting during the trip.
- (a) Verify that $U(n)$ is a Markov chain, and write down its transition matrix.
- (b) Show that the chain is reversible in equilibrium with a stationary distribution such that $\pi_0 = q/(m + q)$.

- (c) Deduce that the expected number w of trips between occasions when I must set off in the rain is $w = (m + q)/(pq)$. Show that w takes its smallest possible value s when

$$p = m + 1 - [m(m + 1)]^{\frac{1}{2}}$$

and that in this case $s = \{2m + 1 - 2[m(m + 1)]^{\frac{1}{2}}\}^{-1}$

$$= 4m + 2 - 1/(4m) + 1/(8m^2) + o(m^{-2}) \text{ as } m \text{ increases.}$$

- (d) When I own only one umbrella, and I have it to hand at the start of the first trip, show that the probability generating function of the number X of trips until I get wet is

$$\mathbf{E}\theta^X = \frac{pq\theta^2}{1 - p\theta + q^2\theta^2}.$$

Find $\mathbf{E}X$, and s and the corresponding value of p in this case.

Appendix: Solutions and Hints for Selected Exercises and Problems

No experienced mathematician feels well acquainted with a subject until he has tackled some problems; through attempting and failing, we extend the boundaries of our knowledge and experience. This observation applies to students also. *It would be a big mistake to treat the remarks of this section as a solution sheet.* Many of the hints and comments will be useful only to those who have spent a half hour, say, on the problem already. The remarks vary in style and content between small hints and detailed solutions; some problems receive no comments at all (indicating, perhaps, that they are either very easy or good challenges).

CHAPTER 1

Exercises

- 1.8.1** $\mathbf{P}(\{1, 1\} \cup \{1, 2\} \cup \{2, 1\} \cup \{6, 6\}) = \frac{4}{36} = \frac{1}{9}$.
- 1.8.2** (a) $\frac{1}{2}$; (b) $\frac{1}{2}$; (c) $\frac{1}{4}$.
- 1.8.3** $\frac{11}{18}$
- 1.8.4** $\frac{5}{12}$
- 1.8.5** $\frac{1}{3} - \frac{1}{18}r$
- 1.9.1** $\Omega = \{(i, j) : 1 \leq i < j \leq 2n\}$ and so $|\Omega| = n(2n - 1)$. Likewise, $|\{HH\}| = \frac{1}{2}n(n - 1)$. Hence, $\mathbf{P}(HH) = (n - 1)/(2(2n - 1))$.
- 1.9.2** (a) $\frac{1}{2}$; (b) $\frac{1}{2}$.
- 1.9.3** $\frac{1}{2}$
- 1.9.4** $\frac{1}{2}$
- 1.10.1** (a) zero; (b) $\frac{2}{3}$.
- 1.10.2** Let C_j be the event that the j th cup and saucer match. Then $\mathbf{P}(C_j) = \frac{1}{4}$; $\mathbf{P}(C_i \cap C_j) = \frac{1}{12}$, $i \neq j$, $\mathbf{P}(C_i \cap C_j \cap C_k) = \frac{1}{24}$, $i \neq j \neq k \neq i$.

Hence, by (1.4.5), (1.4.8), (1.6.1), $\mathbf{P}\left(\bigcap_{i=1}^4 C_i^c\right) = 1 - 4 \cdot \frac{1}{4} + 6 \cdot \frac{1}{12} - 4 \cdot \frac{1}{24} + \frac{1}{24} = \frac{3}{8}$.

1.11.1 $\mathbf{P}(A) = \frac{36}{91}; \mathbf{P}(B) = \frac{30}{91}; \mathbf{P}(C) = \frac{25}{91}$.

1.11.2 Let all the players continue rolling, even after first rolling a 6, and let H_r be the event that all three roll a 6 in the r th round. Because $E \subseteq \left(\bigcup_{r=1}^{\infty} H_r\right)^c$, we have

$$\mathbf{P}(E) \leq 1 - \mathbf{P}\left(\bigcup_{r=1}^{\infty} H_r\right) = 1 - \sum_{r=1}^{\infty} \mathbf{P}(H_r) = 1 - \sum_{r=1}^{\infty} (6^3 - 1)^{r-1} / 6^{3r} = 0.$$

1.11.3 Ignoring Chryseis, $\mathbf{P}(A) = \sum_{n=0}^{\infty} 5^{2n} / 6^{2n+1} = \frac{6}{11}$.

1.11.4 As above, let all players continue rolling irrespective of 6s achieved. A gets his first 6 on the $(3r + 1)$ th roll in 5^r ways; B and C both have at least one already in $(6^r - 5^r)^2$ ways. Hence, $\mathbf{P}(A \text{ last}) = \sum_{r=1}^{\infty} 5^r (6^r - 5^r)^2 / 6^{3r+1} = \frac{305}{1001}$.

1.12.2 $\mathbf{P}(B_1) = \frac{5}{8}; \mathbf{P}(C) = \frac{1}{4}$.

1.12.3 (a) $\mathbf{P}(B_2) = \frac{3}{8}; \mathbf{P}(B_3) = \frac{1}{8}$. (b) $\mathbf{P}(B_2) = \frac{1}{8}; \mathbf{P}(B_3) = \frac{1}{8}$.

(c) $\mathbf{P}(B_2) = \frac{1}{8}; \mathbf{P}(B_3) = \frac{1}{8}$.

1.12.4 (a) 0; (b) $\frac{1}{4}$; (c) $\frac{1}{2}$.

1.13.1 $p_2 = \frac{n_7}{36} + \frac{n_{11}}{36} + \frac{1}{(36)^2} (n_4^2 + n_5^2 + n_6^2 + n_8^2 + n_9^2 + n_{10}^2) = \frac{97}{324}$.

1.13.2 $p_3 = p_2 + \frac{2}{(36)^3} (27n_4^2 + 26n_5^2 + 25n_6^2) = \frac{97}{324} + \frac{1}{6} \cdot \frac{107}{324}$.

1.13.3 (a) $\frac{436}{990}$; (b) $\frac{526}{990}$.

1.13.4 Let p_j be the chance of winning when the first die shows j . Then

$$p_1 = \frac{361}{990}, p_3 = p_4 = \frac{502}{990}, p_5 = \frac{601}{990}, \text{ so you would fix the first die at 5 if you could.}$$

1.14.3 $\sum_{r=0}^{\infty} \left(\frac{1}{2}\right)^{2r} \frac{1}{2} = \frac{2}{3}$.

1.14.4 Let s be a sequence of length $m + n$ in which the first m terms have x heads and $m - x$ tails, and the next n have x heads and $n - x$ tails. Now change the first x heads to tails and the first $m - x$ tails to heads, giving a sequence t of length $m + n$ with n heads. This map is 1-1, so the number of t -sequences equals the number of s -sequences, giving the result.

Problems

1 $\frac{1}{13}$

2 $\frac{1}{13}$

3 (a) $\frac{8}{15} \left(> \frac{1}{2}\right)$; (b) $\frac{112}{225} \left(< \frac{1}{2}\right)$.

- 4 $\frac{1}{2}$
- 5 2^n for some integer $n \geq 1$.
- 6 (a) $\frac{25}{216}$; (b) $\frac{125}{216}$; (c) 1; (d) 17 because $\left(\frac{5}{6}\right)^{17} < 0.05 < \left(\frac{5}{6}\right)^{16}$.
- 7 (a) $(A \cap B) \cup (B \cap C) \cup (C \cap A)$;
 (b) $(A^c \cap B \cap C) \cup (A \cap B^c \cap C) \cup (A \cap B \cap C^c)$; (c) $(A \cap B \cap C)^c$
- 8 The loaded die is equivalent to a fair 10-sided die with five faces numbered 6.
 (a) $\left(\frac{5}{10}\right)^4 = 81 \left(\frac{1}{6}\right)^4$ so the factor is 81; (b) $p_{23} = 4p_{24}$.
- 9 (a) $\frac{5}{6}$; (b) $\frac{1}{4}$; (c) $\frac{3}{16}$; (d) $\frac{1}{8}$.
- 10 $\left(\frac{35}{36}\right)^{24} \simeq 0.16$
- 11 $\frac{19}{36}$
- 12 (a) Use induction; (b) $\frac{1}{9}$. (*Hint: The cups can be arranged in 90 distinct ways.*)
- 13 $\frac{2}{3}$
- 14 $\frac{1}{3}$ in each case.
- 15 (a) $1 - \left(\frac{5}{6}\right)^4$; (b) $\left(\frac{3}{6}\right)^4 - \left(\frac{2}{6}\right)^4$.
- 16 (a) $\left(\frac{1}{2}\right)^n - \left(\frac{1}{3}\right)^n \downarrow 0$; (b) $1 - \left(\frac{5}{6}\right)^n \uparrow 1$.
- 17 (a) $\frac{1}{n} \left[\frac{n}{3} \right]$; $\frac{1}{n} \left[\frac{n}{4} \right]$; $\frac{1}{n} \left[\frac{n}{12} \right]$; $\frac{1}{n} \left(\left[\frac{n}{3} \right] + \left[\frac{n}{4} \right] - \left[\frac{n}{12} \right] \right)$; (b) $\frac{1}{3}$; $\frac{1}{4}$; $\frac{1}{12}$; $\frac{1}{2}$.
 (c) You would get the same answers.
- 19 (c) Use induction.
- 21 In every case, $p = \frac{x(x-1)}{(x+y)(x+y-1)}$ (a) $x = 3, y = 1$ and $x = 15, y = 6$;
 (b) $x = 6, y = 10$; (c) when $r = 6, x = 2$ and $y = 7$.
- 22 Always in each case, except (c), which holds when $B \subseteq C \subseteq A$, and (d), which holds when $A \cap C = \phi$.
- 23 1988 was a leap year, so $p = 1 - \frac{(366)!}{(366-m)!(366)^m}$

CHAPTER 2

Exercises

2.6.6 $1 - r^2$

2.6.7 $\mathbf{P}(V|A_n \cup B_n) = \mathbf{P}(A_n)/\mathbf{P}(A_n \cup B_n) = \frac{r^{n-1}p}{r^{n-1}p + r^{n-1}q} = \frac{p}{p+q}$. (Note the lack of any dependence on n .)

2.7.1 Use induction.

2.7.2 $(c+d)/(b+c+d)$

2.7.3 $(c+nd)/(b+c+nd) \rightarrow 1$ as $n \rightarrow \infty$

2.7.4 Use (2.6.1).

2.7.5 It is also the probability of getting m cyan balls and n blue balls in any given fixed order.

2.8.4 $(1 - p)^n; np(1 - p)^{n-1}$

2.8.5 (a) $1 - \pi_n$; (b) $\sum_{k=1}^{n-1} p_k \pi_{k-1} s_{n-k}$; (c) $1 - (1 - p)^n$; $\frac{1}{2}n(n-1)p^2(1 - p)^{n-2}$.

2.8.7 $\prod_{k=1}^{\infty} (1 - p_k) > 0$, if and only if $\sum_{k=1}^{\infty} p_k < \infty$.

2.9.1 $\alpha(1 - \alpha) + \alpha^2(1 - \gamma)/(1 - (1 - \alpha)(1 - \gamma))$

2.9.2 $(2 - \alpha + \alpha\gamma(1 - \gamma))^{-1}$

2.9.3 Biggles

2.10.1 (a) $\mathbf{P}(E) = 0.108$; $\mathbf{P}(A^c|D) = \frac{45}{86}$.

(b) $\mathbf{P}(E) = 0.059$; $\mathbf{P}(A^c|D) = \frac{5}{451}$.

2.10.2 $\frac{p(\mu(1 - \pi) + v(1 - \mu))}{p(\mu(1 - \pi) + v(1 - \mu)) + (1 - p)(1 - \pi)}$ (a) $\mathbf{P}(A|D^c) = \frac{1}{46} < \frac{45}{86}$;

(b) $\mathbf{P}(A|D^c) = \frac{6}{61} > \frac{5}{451}$

2.10.3 $\mathbf{P}(L) = (1 - \rho)\mathbf{P}(E) + \rho(1 - \mathbf{P}(E))$;

$$\mathbf{P}(A^c|M) = \frac{(\pi(1 - \rho) + \rho(1 - \pi))(1 - p)}{\rho\mathbf{P}(D^c) + (1 - \rho)\mathbf{P}(D)}$$

$$\mathbf{P}(A|M^c) = \frac{p[(1 - \rho)(\mu(1 - \pi) + v(1 - \mu)) + \rho(\mu\pi + (1 - \mu)(1 - v))]}{\rho\mathbf{P}(D) + (1 - \rho)\mathbf{P}(D^c)}$$

2.11.4 $1 - k/K$

2.11.6 $\frac{(1 - \rho)(1 - \rho^K)}{(1 - \rho^{k+1})(1 - \rho^{K-k})}$, where $\rho = \frac{1 - p}{p}$.

2.12.1 When $\lambda = \mu$.

2.12.2 $\mathbf{P}(A_2|A_1) - \mathbf{P}(A_2) = \frac{(\mu - \lambda)^2}{2(\lambda + \mu)} = 0$. When $\lambda = \mu$.

2.12.3 (a) $(\mu^3 + \lambda^3)/(\mu^2 + \lambda^2)$ (b) $\frac{\mu^n + \lambda^n}{\mu^{n-1} + \lambda^{n-1}} \rightarrow \max\{\lambda, \mu\}$, as $n \rightarrow \infty$

2.12.4 (a) $\frac{\mu}{\lambda + \mu}$; (b) $\frac{\lambda}{\lambda + \mu}$

2.12.5 $\frac{\mu^n}{\mu^n + \lambda^n} \rightarrow 1$, as $n \rightarrow \infty$, if $\mu > \lambda$.

2.13.3 (a) $\frac{2p_1}{3p_1 + p_2(1 - p_1)}$; (b) $\frac{3p_1}{3p_1 + p_2(1 - p_1)}$.

2.13.4 Yes, if $p_1 < \frac{1}{2}$ and $p_2 = \frac{p_1}{1 - p_1}$.

2.14.1 (a) $a = 1, b = 1, c = 2$; (b) $a = b = c$.

2.14.1 $\frac{16}{41}$

2.14.2 $\frac{20}{41}$

Problems

- 1 (a) 0.12; (b) 0.61; (c) 0.4758; (d) $\frac{7}{13}$.
- 3 (i) $\frac{1}{6}; \frac{1}{3}; \frac{1}{2}$ respectively; (ii) $\frac{7}{9}$; (iii) $\frac{1}{14}; \frac{2}{7}; \frac{9}{14}$, respectively; (iv) $\frac{6}{7}$.
- 4 (b) No.
- 5 (a) 0.36; (b) 0.06; (c) 0.7; (d) $\frac{18}{41}$
- 6 (a) zero; (b) one.
- 7 (a) $\frac{3}{5}$; (b) $\frac{2}{3}$; (c) $\frac{5}{6}$; (d) $\frac{4}{5}$.
- 8 (a) $\frac{1}{2}$; (b)(i) $\frac{1}{2}$; (ii) $\frac{1}{4}$; (c) $\frac{1}{36}$; (d) $\frac{1}{42}$.
- 9 (b) $\frac{3}{4}$; (c) 0.7.
- 11 $\frac{83}{102}$
- 12 (a) $(1 - p^2)^2$; (b) $1 - p + p(1 - p^2)^2$; (c) $\frac{p(1 - p^2)^2}{1 - p + p(1 - p^2)^2}$
- 13 (a) $\frac{r}{b+r}$; (b) $\frac{1}{2}$; (c) zero if $r < b$, $\frac{1}{2}$ if $r = b$, 1 if $r > b$.
- 14 $1 - x \leq \exp(-x)$.
- 15 Let $\mathbf{P}(A) = \alpha$ and $\mathbf{P}(B) = \beta$. If the claim is false, then $(1 - \alpha)(1 - \beta) < \frac{4}{9}$ and $\alpha\beta < \frac{4}{9}$ and $\alpha(1 - \beta) + \beta(1 - \alpha) < \frac{4}{9}$. The intersection of these three regions in the $\alpha - \beta$ plane is empty, so the claim is true.
- 16 $\mathbf{P}(E|A = \text{tail}) = (1 - (1 - p)^{s-1})\mathbf{P}(E|A = \text{head})\mathbf{P}(E) = \frac{p^{r-1}(1 - (1 - p)^s)}{1 - (1 - p^{r-1})(1 - (1 - p)^{s-1})}$
- 17 He must answer k satisfying

$$\left(\frac{u(1-b)}{b(1-u)}\right)^k > \frac{p}{q} \left(\frac{1-b}{1-u}\right)^n.$$

Therefore, the examiners must set questions such that $(u/b)^n > p/q$ or the student can never convince them that he is not bluffing.

- 18 (i) 0.7; (ii) $\frac{2}{9}$; (iii) Use (2.11.2).
- 19 The second set of rules.
- 20 $\frac{16}{31}$
- 21 6^{-2}
- 22 (a) $\frac{1}{2} \left(1 - \left(\frac{2}{3}\right)^n\right)$; (b) $\frac{1}{36(\lambda - \mu)} \left(\frac{\lambda^n}{1 - \lambda} - \frac{\mu^n}{1 - \mu}\right)$ where $\lambda = \frac{5}{12} + \frac{\sqrt{5}}{4}$
and $\mu = \frac{5}{12} - \frac{\sqrt{5}}{4}$; (c) $\frac{1}{2}$.

$$23 \quad \mathbf{P}(\text{ever hit}) \leq \frac{1}{4} \sum \left(\frac{2}{3}\right)^{n-1} = \frac{3}{4}.$$

$$24 \quad \text{(i) } \frac{7}{8}; \quad \text{(ii) } \frac{83}{128}.$$

$$25 \quad \text{(i) } p_A = p_B = \frac{1 - \frac{1}{2}(\theta + \theta^2)}{2 - \theta + \theta^2}; \quad \text{(ii) } \theta = \frac{1}{10}(\sqrt{41} - 1); \quad \text{(iii) } \frac{2(1 - \theta)}{2 - \theta - \theta^2}.$$

$$26 \quad \text{(a) } \frac{197}{450}; \quad \text{(b) } \frac{77}{225}; \quad \text{(c) } \frac{25}{148}.$$

$$27 \quad \text{(a) } 5^{n-1} \cdot 6^{-n}; \quad \text{(b) } \frac{5}{11}; \quad \text{(c) } \frac{1}{6}.$$

$$28 \quad \text{(a) } 2 \left(\frac{p \left(\frac{q}{p}\right)^{n-1} - q}{\left(\frac{q}{p}\right)^{n-1} - 1} \right); \quad \text{(b) } (p - q) \left(\frac{1 + \left(\frac{q}{p}\right)^n}{1 - \left(\frac{q}{p}\right)^n} \right)$$

$$\text{(c) } \frac{q - p}{\left(\frac{q}{p}\right)^n - 1} + \frac{q \left(\frac{p}{q}\right)^{n-1} - p}{\left(\frac{p}{q}\right)^{n-1} - 1}.$$

$$29 \quad \text{(i) } p_m^k = \frac{1}{2}p_{m-1}^k + \frac{1}{2}p_{m+1}^k; \quad p_0^k = 0, \quad p_n^k = p_{n-1}^{k-1}$$

$$\text{(ii) } p_m^0 = 1 - \frac{m}{n}; \quad \sum_{k=0}^{\infty} p_m^k = 1 - \frac{m}{n} + \frac{m}{n} = 1.$$

$$30 \quad \text{(a) } \sum_{k=1}^6 \frac{1}{6} \left(1 + \left(\frac{q}{p}\right)^{60/k}\right)^{-1} \quad \text{(b) Choose } x = 6.$$

$$31 \quad \text{(a) } \frac{1 - p_1}{3 - p_1}; \quad \text{(b) } \frac{(1 - p_1)^2}{2 + (1 - p_1)^2}; \quad \text{(c) } \frac{1 - p_1}{3 - p_1 - p_2 - p_3}.$$

$$32 \quad \text{(i) } \frac{1}{16}; \quad \text{(ii) } 5 \text{ metres and one step}; \quad \text{(iii) } 5 \text{ metres}.$$

$$33 \quad \text{(a) } \frac{1}{(n!)^2} \quad \text{(b) } p_t = \begin{cases} \frac{t-1}{jk} & \text{for } 2 \leq t \leq j \wedge k \\ \frac{(j \wedge k) - 1}{jk} & \text{for } j \wedge k \leq t \leq j \vee k \\ \frac{j+k-t-1}{jk} & \text{for } j \vee k \leq t \leq j+k-2 \end{cases} \quad \text{(c) } j = k = n!$$

$$34 \quad \text{(i) } 0.3; 0.1545; \quad \text{(ii) For men, } 0.5; 0.6; \text{ for women, } 0.1; 0.11.$$

$$35 \quad \text{(a) } \frac{11}{23}; \quad \text{(b) } \frac{12}{23}.$$

36 No

$$39 \quad p_n = (1 - p) \left(\frac{1 - (q - p)^{n-1}}{1 - (q - p)^n} \right) \rightarrow 1 - p.$$

$$44 \quad \mathbf{P}(\text{win with 1-6 flat}) = 0.534;$$

$$\mathbf{P}(\text{win with 5-2 flat}) = 0.5336.$$

CHAPTER 3

Exercises

- 3.9.1** (a) $\frac{(n+r)!}{r!}$; (b) $\binom{n+r}{r}$; (c) $p_{r+1}(n)$.
- 3.9.2** (i) 462, assuming oranges are indistinguishable.
 (ii) $7 = \underline{x^5}$ (the coefficient of x^5) in $\prod_{k=1}^7 (1-x^k)^{-1}$.
- 3.9.3** 4
- 3.10.1** (a) $\frac{6}{n(n-1)}$; (b) $\frac{6(n-3)}{n(n-1)}$; (c) $\frac{5!(n-5)(n-6)}{2n(n-1)(n-2)(n-3)}$
- 3.11.7** $\frac{1}{n^r} \sum \frac{n!}{k!(r-2k)!(n-r+k)!}$, where the sum is over all k such that $\max\{0, r-n\} \leq 2k \leq r \leq 2n$.
- 3.11.8** $\frac{1}{n^r} \sum_{k=0}^n \binom{n}{k} (-)^k M_k$, where $r \geq 2n$ and $M_k = (n-k)^{r-k} (n-k+1)^k$.
- 3.12.1** $n^{-r} \binom{n}{r}$
- 3.12.3** $p(m, n) - p(m-1, n)$, where $p(m, n)$ is given in Example 3.11.
- 3.13.3** Rotational symmetry.
- 3.13.4** As in the example, k disjoint pairs of seats can be chosen in M_k ways, k given pairs of twins can occupy the pairs of seats in $2^k k!$ ways, the rest occupy their seats in $(2n-2k)!$ ways, so \mathbf{P} (no pair adjacent) =

$$\sum_{k=0}^n (-)^k \binom{n}{k} \frac{M_k 2^k k! (2n-2k)!}{(2n)!} = \sum_{k=0}^n \frac{(-)^k}{k!} \frac{n! (2n-k-1)!}{(n-k)! (2n-1)!} 2^k \rightarrow e^{-1}$$
 as $n \rightarrow \infty$.
- 3.13.5** e^{-2} .
- 3.15.1** $b \binom{a}{k} / \binom{n}{k+1}$
- 3.15.2** Same as 3.15.1
- 3.15.3** $\left(1 + \binom{n-k-1}{a-k+1}\right) / \binom{n}{a}$
- 3.15.4** $\left(b \binom{a}{k} + a \binom{b}{k}\right) / \binom{n}{k+1}$
- 3.16.2** (a) This depends on the order in which you catch the species, so the answer is a horrible sum. (b) $\binom{n}{x} \frac{b^x c^{n-x}}{(b+c)^n}$. Avoid sampling with partial replacement!
- 3.16.4** $\binom{n}{x} p^x (1-p)^{n-x}$ in all three cases.
- 3.17.5** $K(20, 3) = \binom{22}{2} = 231$.
- 3.17.6** $K(17, 3) = \binom{19}{2} = 171$.
- 3.17.7** Recall Example 3.12.
- 3.18.5** $p(n, r+s) / \sum_{j=r}^n p(n, j)$. Let $n \rightarrow \infty$.

$$3.18.7 \quad (a) \frac{1}{n!r!} \sum_{k=0}^{n-r} (-)^k (n-r-k)!/k!; \quad (b) \sum_{k=0}^n (-)^k \frac{(n-k)!}{n!k!}; \quad (c) (p(n, 0))^2.$$

$$3.18.8 \quad (a) \text{ zero}; \quad (b) \text{ zero}; \quad (c) e^{-2}.$$

Problems

$$1 \quad \frac{49}{153}$$

$$2 \quad (a) \frac{6a!b!c!}{(a+b+c)!}; \quad (b) \frac{1}{a!b!c!}; \quad (c) \frac{6}{(a+b+c)!}.$$

$$3 \quad (a) \frac{4!48!}{(12!)^4} \bigg/ \frac{52!}{(13!)^4}; \quad (b) \frac{16}{\binom{52}{13}} - \frac{72}{\binom{52}{13}\binom{39}{13}} + \frac{72}{\binom{52}{13}\binom{39}{13}\binom{26}{13}};$$

$$(c) \left(4\binom{39}{13} - 6\binom{26}{13} + 4\right) \bigg/ \binom{52}{13}; \quad (d) 4\binom{48}{9} \bigg/ \binom{52}{13}.$$

$$4 \quad (a) 13\binom{4}{2}\binom{12}{3}4^3 \bigg/ \binom{52}{5}; \quad (b) 44\binom{4}{2}^2\binom{13}{2} \bigg/ \binom{52}{5};$$

$$(c) 4^5 10 \bigg/ \binom{52}{5}; \quad (d) 4\binom{13}{5} \bigg/ \binom{52}{5}; \quad (e) 156\binom{4}{2}\binom{4}{3} \bigg/ \binom{52}{5}.$$

$$6 \quad \frac{32491}{1+10^5}$$

$$7 \quad (i) 10; \quad (ii) S, \text{ where } S = \max\{n : n^2 \leq M\}.$$

$$8 \quad (a) \binom{m+r}{m} \text{ is an integer.} \quad (b) \text{ You have } (k-1)! \text{ colours, and } k \text{ balls of each colour. How many arrangements are there?}$$

$$9 \quad (a) 5\left(\frac{5}{6}\right)\left(\frac{1}{6}\right)^4 \text{ if the question means exactly 4 aces, or } \frac{26}{6}\left(\frac{1}{6}\right)^4 \text{ if it means at least 4 aces.}$$

$$(b) 6 \text{ times (a).} \quad (c) 1 - 3\left(\frac{5}{6}\right)^5 + 3\left(\frac{2}{3}\right)^5 - \left(\frac{1}{2}\right)^5.$$

$$10 \quad (a) 18 \bigg/ \binom{64}{8}; \quad (b) 8! \bigg/ \binom{64}{8}.$$

$$11 \quad (a) 1 - \binom{3n}{r} \bigg/ \binom{4n}{r}; \quad (b) \binom{n}{2} \binom{3n}{r-2} \bigg/ \binom{4n}{r};$$

$$(c) \left(\binom{4n}{r} - 4\binom{3n}{r} + 6\binom{2n}{r} - 4 \right) \bigg/ \binom{4n}{r}.$$

$$12 \quad (a) 15; \quad (b) 2; \quad (c) 36.$$

$$14 \quad P_k = \binom{b+w-k-1}{w-1} \bigg/ \binom{b+w}{w} = \frac{w}{b+w} \binom{b}{k} \bigg/ \binom{b+w-1}{k}.$$

$$15 \quad (b)(i) \binom{N+m-1}{m-1}; \quad (ii) \binom{N-1}{m-1}.$$

- 16 Follows from $-\frac{1}{2} + \left(Kn + \frac{1}{4}\right)^{\frac{1}{2}} < N < \frac{1}{2} + \left(Kn + \frac{1}{4}\right)^{\frac{1}{2}}$.
- 17 (a) k^n ; (b) $k!$ if $k \leq n$; (c) $\binom{n+k-1}{n}$.
- 19 See Example 3.11.
- 20 $\sum_{m=0}^n \frac{(-)^m}{m!} \left(\frac{n!}{(n-m)!}\right)^2 \frac{(2n-2m)!}{2n!} 2^m \sim \sum_0^{\infty} \frac{(-)^m}{m!} \frac{1}{2^m}$.
- 23 (a) $\frac{(n-4)(n-3)}{(n-2)(n-1)}$ whether Arthur sits at random or not.
(b) Number the knights at the first sitting, and then use Problem 21.
- 25 (a) $\sum_{k=0}^n (-)^k \binom{n}{k} \frac{(2n-k)!}{(2n)!} \rightarrow e^{-\frac{1}{2}}$.
(b) Problem 21 again. The limit is e^{-1} .
- 26 $\frac{1}{2}$; (ii) $7 / \binom{7}{3} = \frac{1}{5}$; (iii) $\frac{1}{3} \binom{7}{2} 2^{-7} + \left(\frac{4}{5}\right) \binom{7}{3} 2^{-7} + \frac{1}{2} = \frac{99}{128}$; (iv) $\frac{92}{99}$.
- 27 $\binom{2n}{n} (pq)^n = (4pq)^n \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{4}\right) \dots \left(1 - \frac{1}{2n}\right)$
 $\leq (4pq)^n \exp\left(-\frac{1}{2} \sum_1^n k^{-1}\right) \leq (4pq)^n \exp\left(-\frac{1}{2} \log n\right) \rightarrow 0$.
- 36 $\frac{1}{4} + \frac{3}{4} \left(\frac{-1}{3}\right)^n$.
- 37 The number of ways of choosing k nonoverlapping triples (three adjacent) is $\binom{3n-2k}{k}$. Now use inclusion-exclusion.
- 39 Use induction.

CHAPTER 4

Exercises

- 4.8.2 (a) 992; (b) $32k!(31)^{k-1}$.
- 4.8.3 They give the same chance because trials are independent.
- 4.8.4 $\mathbf{P}(T > j + k | T > j) = (31/32)^k = \mathbf{P}(T > k)$.
- 4.9.4 Either form a difference equation or rearrange the sum.
- 4.9.6 (a) $[(n+1)p - 1]$; (b) $\left[\frac{k}{p} - 1\right]$.
- 4.9.7 (a) $\frac{1}{2}(1 + (2p-1)^n)$; (b) $\frac{1}{2}(1 - (2p-1)^n)$;
 (c) $\sum_{m=k}^n \frac{m!}{(m-k)!} \binom{n}{m} p^m (1-p)^{n-m} = \frac{p^k n!}{(n-k)!}$; $0 \leq k \leq n$.
- 4.10.1 $(1 - e^{-\lambda} - \lambda e^{-\lambda}) / (1 - e^{-\lambda})$
- 4.10.2 $\lambda(1 - e^{-\lambda})^{-1}$
- 4.10.4 (a) $[\lambda - 1]$; (b) k .

4.10.5 (a) $\exp(\lambda(e-1))$; (b) $e^{-\lambda}$; (c) λ ; (d) λ^k .

4.11.3 Choose t to minimize

$$L(t) = a \sum_{x \leq t} (t-x)f(x) + b\mathbf{P}(X > t).$$

$$(a) \hat{t} = \left\lceil \frac{\log a - \log(a+bp)}{\log q} \right\rceil + 1; \quad (b) \hat{t} = \left\lceil \left(1 + \frac{b}{a}\right)^{\frac{1}{2}} \right\rceil;$$

$$(c) \hat{t} = \left\lceil \frac{b}{a} - n \right\rceil \text{ if this lies in } [-n, n]; \text{ otherwise, } \hat{t} = n \text{ if } b/a > 2n.$$

(What if a or b can be negative?).

4.11.4 (a) Any median of X ; (b) $\mathbf{E}(X)$.

4.11.5 Minimize $L(m) = b \sum_{k=0}^m (m-k)p_k(m) + c \sum_{k=m+1}^{m+n} (k-m)p_k(m)$, where

$$p_k(m) = p^k(1-p)^{m+n-k} \binom{m+n}{k}. \text{ For no overbooking, you need}$$

$L(0) < L(m)$ for all $m > 0$; solutions are approximate or numerical.

4.12.2 $k(a+b-k)$.

4.12.3 Let p_k be the probability that B wins if A 's initial fortune is k . Then $\mathbf{P}(A|B) = \mathbf{P}(B|A)\mathbf{P}(A)/\mathbf{P}(B) = pp_{k+1}/p_k$ and $\mathbf{P}(A^c|B) = (1-p)p_{k-1}/p_k$.

Hence, $pp_{k+1}\mathbf{E}(X_{k+1}|B) - p_k\mathbf{E}(X_k|B) + (1-p)p_{k-1}\mathbf{E}(X_{k-1}|B) = -p_k$.

When $p = \frac{1}{2}$, we have $p_k = (a+b-k)/(a+b)$, giving

$$\mathbf{E}(X_k|B) = \frac{1}{3}((a+b)^2 - (a+b-k)^2), \quad 0 \leq k < a+b.$$

$$4.12.4 \sum_{k=0}^{a+b} m_k \binom{a+b}{k} 2^{-(a+b)} = \begin{cases} \frac{a+b}{q-p} \left\{ \frac{1}{2} - \frac{1 - (2p)^{-(a+b)}}{1 - (q/p)^{a+b}} \right\}; & p \neq q \\ \frac{1}{4}(a+b)(a+b-1); & p = \frac{1}{2}. \end{cases}$$

$$4.12.5 (i) p \neq q; \frac{k - (a+b)}{q-p} + \frac{p}{(q-p)^2} \left(\left(\frac{q}{p}\right)^{a+b} - \left(\frac{q}{p}\right)^k \right).$$

$$(ii) p = \frac{1}{2}; (a+b)(a+b-1) - k(k-1).$$

$$4.13.1 (b) \frac{M+1}{j(j+1)M}; \quad (c) \frac{1 - \exp(-\lambda j)}{1 - \exp\left(-\frac{1}{2}\lambda M(M+1)\right)}.$$

4.13.2 $\mathbf{E}(X_A) = 2$.

$$4.13.3 (b) \sum_{n=1}^{\infty} \frac{m+1}{m+n} = \infty; \quad (c) e^{-\frac{1}{2}m(m+1)} \sum_{n=1}^{\infty} e^{-\frac{1}{2}\lambda(m+n)(m+n+1)} < \infty.$$

4.13.4 (i) Median is ∞ ; (ii) $\mathbf{E}(X_A|X_A < \infty) = \infty$.

$$4.14.1 \hat{m} = \max \left\{ 0, \left\lceil \frac{N(b+d) - c}{b+d+c} \right\rceil \right\}.$$

4.14.2 $\hat{m} = [m]$, where m is the positive root of $(m+1)q^m = c/(b+c)$.

4.15.1 You win \$1 with probability 1, but your winning bet has infinite expected value.

4.15.2 $\frac{p(1-(2q)^{L+1})}{(1-2q)(1-q^{L+1})}$ if $q \neq p$. This $\rightarrow \infty$ if $p < \frac{1}{2}$ or $\rightarrow p/(2p-1)$ if $p > \frac{1}{2}$.

If $p = \frac{1}{2}$, then expectation is $\frac{Lp}{1-q^{L+1}} \rightarrow \infty$ as $L \rightarrow \infty$.

4.16.2 With an obvious notation $m_{us} = \frac{2}{3} \cdot \frac{1}{4} \cdot \left(\frac{3}{4}\right)^{s-1} = 2 \cdot 3^{s-2} 4^{-s}$;

$m_{ds} = \frac{1}{3} \cdot \frac{3}{4} \cdot \left(\frac{1}{4}\right)^{s-1} = 4^{-s}$. Hence, the r th search downstairs comes after the s th search upstairs if $\frac{2}{9} > 3^{-s} 4^{s-r} > \frac{1}{6}$.

The order is *duuuuuuuuuudu...*

4.16.3 Place $m_{rs} = (1-d_{r1})(1-d_{r2}) \dots (1-d_{rs})p_r$ in nonincreasing order.

4.17.4 $q^2(1+p)/(1-pq) = (1-p^2)^2/(1+p^3)$

4.17.5 $(1-q^2)^2/(1+q^3)$

4.17.6 $\mathbf{E}(X|B^c) = \frac{2}{1-pq} + \frac{q}{1+q}$

4.17.7 $\mathbf{P}(A_1|B) = \frac{p}{1+p}$; $\mathbf{P}(A_2|B) = \frac{qp}{1+p}$.

4.17.8 Every number in $[2, 3)$ is a median.

4.17.9 $\mathbf{P}(B) = p^2(1-q^3)/(1-(1-p^2)(1-q^2))$

$\mathbf{E}(X) = (1+pq)(1-2pq)/(1-pq(1-p)(1+q))$

4.17.10 With new rules $\mathbf{P}(B) = \frac{q^2}{p^2+q^2}$, which is smaller than old $\mathbf{P}(B)$ if $p > \frac{1}{2}$.

Brianchon is making a mistake. $\mathbf{E}(X) = 4$ with new rules.

4.18.2 Because the answer "yes" is false with probability $(1-p)/(2p)$, individuals should be much more likely to tell the truth. Then $\frac{2Y_n}{n} - 1$ should not be too far from p in the long run.

4.18.3 $\mathbf{P}\left(\bigcup_n A_k(\epsilon)\right) \leq \sum_n \mathbf{P}(A_k(\epsilon))$.

4.18.4 Use 4.18.3

4.18.5 Use 4.18.1

4.18.6 Use Markov's inequality and Chebyshev's inequality

4.18.8 $a(m, k) = \frac{m(m-1)\dots(m-k+1)}{(m+1)\dots(m+k)}$. Hence, for large enough m with k fixed

$$\left| \log a(m, k) + \frac{1}{m} + \frac{2}{m} + \dots + \frac{k-1}{m} + \frac{1}{m} \dots + \frac{k}{m} \right| \leq \frac{2 \sum_1^k r^2}{m^2} \rightarrow 0 \text{ as } m \rightarrow \infty.$$

Hence, $[a(m, k)]^m e^{k^2} \rightarrow 1$. The inequalities follow from

$$\binom{2m}{m} 4^{-m} < \frac{2m(2m-2)\dots 2}{(2m+1)\dots 3 \cdot 1} = \frac{1}{2m+1} \left(\binom{2m}{m} 4^{-m} \right)^{-1}$$

and

$$2 \binom{2m}{m} 4^{-m} > \frac{(2m-2)\dots 4 \cdot 2}{(2m-1)\dots 3 \cdot 1} = \frac{1}{2m} \left(\binom{2m}{m} 4^{-m} \right)^{-1}.$$

4.19.6 For the left inequality, prove and use the fact that for any collection of probabilities

p_1, \dots, p_r , we have $-\sum_i p_i \log \left(\sum_i p_i \right) < -\sum_i p_i \log p_i$. Equality holds when $g(\cdot)$ is a one-one map. For the right-hand inequality, note that

$$f_i = \exp(-cg(x_i)) / \sum_i \exp(-cg(x_i))$$

is a mass function and use 4.19.1. Equality holds if $f_i = f_X(x_i)$ for all i .

Problems

1 (b) $f(0) = \frac{1}{140}$, $f(1) = \frac{18}{140}$, $f(2) = \frac{66}{140}$; $f(3) = \frac{55}{140}$, $\mathbf{E}(X) = \frac{9}{4}$.

3 $\frac{35}{12}$

4 If X is uniform on $\{1, 2, \dots, n\}$, then $\text{var}(X) = \frac{1}{12}(n^2 - 1)$.

5 (a)(i) $(e^2 - 1)^{-1}$; (ii) $p^{-1} - 1$; (iii) $(\log(1 - p)^{-1})^{-1}$; (iv) $6\pi^{-2}$; (v) 1.

(b)(i) $2e^2(e^2 - 1)^{-1}$; (ii) $(1 - p)^{-1}$; (iii) $p((1 - p) \log(1 - p)^{-1})^{-1}$; (iv) ∞ ; (v) ∞ .

6 Yes, in all cases.

7 $c = \frac{4(M+1)(M+2)}{M(M+3)} \rightarrow 4$; $\mathbf{E}(X) = \frac{2(M+1)}{M+3} \rightarrow 2$.

8 Condition on the appearance of the first tail to get

$$\mathbf{P}(A_n) = \frac{1}{2}\mathbf{P}(A_{n-1}) + \frac{1}{4}\mathbf{P}(A_{n-2}) + \frac{1}{8}\mathbf{P}(A_{n-3}), \quad n > 3.$$

Hence, $\mathbf{P}(A_n) = A\alpha^n + B\beta^n + C\gamma^n$,

where α, β, γ are roots of $8x^3 - 4x^2 - 2x - 1 = 0$, and A, B, C are chosen to ensure that

$$\mathbf{P}(A_1) = \mathbf{P}(A_2) = 0 \text{ and } \mathbf{P}(A_3) = \frac{1}{8}.$$

Similar conditioning gives

$$\mathbf{E}(T) = \frac{1}{2}(1 + \mathbf{E}(T)) + \frac{1}{4}(2 + \mathbf{E}(T)) + \frac{1}{8}(3 + \mathbf{E}(T)) + \frac{3}{8}.$$

Hence, $\mathbf{E}(T) = 14$. To find $\mathbf{E}(U)$, consider the event that a sequence of n tosses including no HTH is followed by HTH . Hence, either $U = n + 1$ or $U = n + 3$, and so

$$\mathbf{P}(U > n) \frac{1}{8} = \mathbf{P}(U = n + 1) \frac{1}{4} + \mathbf{P}(U = n + 3).$$

Summing over n gives $\frac{1}{8}\mathbf{E}(U) = \frac{1}{4} + 1$; $\mathbf{E}(U) = 10$.

9 $\mathbf{E}(X) = \infty$.

14 (i) $\frac{1}{b-a+1}$ for $a \leq k \leq b$; (ii) $\frac{m-n-k}{m-n}$.

15 e.g. $f(-2) = \frac{1}{2}$, $f(1) = f(3) = \frac{1}{4}$.

16 (i) $\sum_{-\infty}^{\infty} (f(2n) - f(2n+1))$; (ii) zero.

$$17 \quad F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right); & a > 0 \\ \mathbf{P}\left(X \geq \frac{y-b}{a}\right); & a < 0 \\ \begin{cases} 0 & y < b \\ 1 & y \geq b \end{cases}; & a = 0. \end{cases}$$

19 (a) When $\lambda = 0$; (b) $p^{-1} - \log(1-p)^{-1} - 1$.

20 (a) If the type is uncommon and the population is large. (c) X is roughly Poisson with parameter 10 so \mathbf{P} (this is the only one) $\simeq 10(e^{10} - 1)^{-1}$, which is very small.

$$21 \quad f_R(r+1) = \binom{n-r}{n}^2, \quad f_R(r-1) = \binom{r}{n}^2; \quad \mathbf{E}(R) \rightarrow \frac{1}{2}n.$$

$$22 \quad f_X(1) = \frac{5}{18}; \quad f_X(3) = \frac{2}{7}; \quad \mathbf{E}(X) = 2 - \frac{5}{18} + \frac{2}{7}.$$

$$24 \quad \mathbf{P}(X \leq n) = \frac{1}{2}(m+n)(m+n+1)m^{-2} \text{ for } -m \leq n \leq 0. \text{ For}$$

$$0 \leq n \leq m, \quad \mathbf{P}(X \leq n) = 1 - \frac{1}{2}(m-n)(m-n-1)m^{-2}.$$

25 Mean μp .

$$26 \quad \text{(i) } \frac{6!}{(2!)^3}; \quad \text{(ii) } \frac{6}{5} + \frac{16}{15} + \frac{5}{3} = \frac{59}{15}.$$

$$27 \quad \text{(a) } p^k q^{n-k} \binom{n-1}{k-1}; \quad \text{(b) } (rp)^k (1-rp)^{n-k} \binom{n-1}{k-1};$$

$$\text{(c) } (1-r)^k p^k (1-p)^{n-k} \binom{n}{k} (1-pr)^{-n}; \quad \mathbf{E}(X) = \frac{n(1-r)p}{1-rp}.$$

28 (i) Choose A if $2000(1-p) < 1000p$ (i.e., if $p > \frac{2}{3}$).

(ii) Choose A if $2000\left(1 - \frac{9p}{4+5p}\right) < 1000\left(\frac{9p}{4+5p}\right)$, but $p > \frac{8}{17}$, so choose B .

29 (a) p^n ; (b) $(1-p)p^{n-1}$; (c) $(1-p)^{-1}$.

$$31 \quad M(n)/n \rightarrow C = \frac{1}{1+p} \text{ as } n \rightarrow \infty.$$

32 (a) $1 - (1-p)^7$; (b) $p(1-p)^{x-1}/(1 - (1-p)^7)$;

$$\text{(c) } \sum_1^7 x f_X(x) = \frac{1}{p} - \frac{7(1-p)^7}{1 - (1-p)^7}; \quad \text{(d) } \frac{1}{p}(1 - (1-p)^4).$$

36 (b) $10^2(1-p)^{10} + 10^3 p(1-p)^9$; (c) $[10^2(1-p)^{10} + 990p(1-p)^{19}]b + (1-r)10^2c$.

37 (c) $\frac{2}{3}(n+1)$.

38 (a) Y is $B\left(n, \frac{1}{36}\right)$; $\mathbf{E}(Y) = n/36$. (b) $X-r$ is $B\left(n-r, \frac{6}{7}\right)$; $\mathbf{E}(X-r) = (n-r)\frac{6}{7}$.

39 (a) $\left(\frac{p^2}{1+p} + \frac{q^2}{1+q}\right)^{-1}$; $\frac{p^2}{1+p} \left(\frac{p^2}{1+p} + \frac{q^2}{1+q}\right)^{-1}$.

$$\text{(b) } \frac{1}{1-r} \left(\frac{2(1-r)^2 + pq}{(1-r)^2 - pq}\right); \quad \frac{p^2}{p+q} \left(\frac{p^2 + 2q}{p^2 + q^2 + pq}\right).$$

The second set of rules.

41 $B\left(n, \frac{1}{32}\right)$, use the Poisson approximation.

$$\begin{aligned} 43 \quad n - \mathbf{E}(H) &= \mathbf{E}(n - H) = \sum_{k=0}^n (n - k)p_k = \sum_0^{n-1} (2n - k) \binom{2n - k - 1}{n} 2^{k-2n}. \\ &= \sum_1^n (2n - k + 1) \binom{2n - k}{n} 2^{k-1-2n} \\ &= \frac{1}{2} \mathbf{E}(2n + 1 - H) - \frac{2n + 1}{2^{2n+1}} \binom{2n}{n}. \end{aligned}$$

$$\text{So } \mathbf{E}(H) = \frac{2n + 1}{2^{2n}} \binom{2n}{n} - 1.$$

$$44 \quad \sum_k q(1 - q)^{k-1} a_k$$

45 With an obvious notation, $m = 1 + \frac{1}{2}m_1 + \frac{1}{3}m_2 + \frac{1}{6}m_3$, also $m_1 = 1 + \frac{1}{2}m_1 + \frac{1}{3}m_{12} + \frac{1}{6}m_{13}$ (and two similar equations), also $m_{12} = 1 + \frac{5}{6}m_{12}$ (and two similar equations). Solve to get m .

CHAPTER 5

Exercises

$$5.11.3 \quad \rho = - \left(\frac{qr}{(1-q)(1-r)} \right)^{\frac{1}{2}}$$

$$5.11.4 \quad \binom{m}{x} \left(\frac{p}{p+q} \right)^x \left(\frac{q}{p+q} \right)^{m-x}$$

$$5.11.5 \quad \sum_{k=0}^9 \frac{18!(pq)^{9-k} r^{2k}}{((9-k)!)^2 (2k)!}$$

$$5.11.6 \quad \frac{p}{p+r} (n - Y)$$

$$5.11.7 \quad n - Y - Z$$

5.12.1 S is $B(m, p^2)$ with mean mp^2 , variance $mp^2(1 - p^2)$.

5.12.5 A is the sum of two binomial random variables $B(m, \phi)$ and $B(m, \mu)$. Hence,

$$f_A(a) = \sum_{k=0}^a \phi^k (1 - \phi)^{m-k} \binom{m}{k} \mu^{a-k} (1 - \mu)^{m-a+k} \binom{m}{a-k}$$

Therefore $\mathbf{E}(A) = m(\phi + \mu)$.

$$5.12.6 \quad \mathbf{E}(A|S) = 2S + (m - S)\phi + (m - S)\mu$$

$$5.13.3 \quad \mathbf{E}(R_n) = 2 + \frac{2}{3} \frac{2^n - 1}{2^{n-1} + 1} + \frac{1}{3} \frac{2^{n-1} - 1}{2^n + 1} \rightarrow \frac{7}{2}$$

$$5.13.4 \quad p_r = \frac{r}{r+b}$$

$$5.13.5 \quad \frac{m}{n}$$

5.14.4 When $X_i = c$, where c is constant.

5.14.5 We assume that "at random" means an individual is selected at random from n independent families X_1, \dots, X_n . Define

$$I(X_i \geq k) = \begin{cases} 1 & \text{if } X_i \geq k \\ 0 & \text{if } X_i < k. \end{cases}$$

Then

$$f_R(k) = \mathbf{E} \left(\frac{\sum_{i=1}^n I(X_i \geq k)}{\sum_{i=1}^n X_i} \right) = n \mathbf{E} \left(\frac{I(X_1 \geq k)}{\sum_{i=1}^n X_i} \right)$$

and

$$\mathbf{E}(R) = \mathbf{E} \left(\frac{\sum_k k I(X_1 \geq k)}{\sum_1^n X_i/n} \right).$$

5.14.6 Let X be uniform on $\{x_1, \dots, x_n\}$ and Y uniform on $\{y_1, \dots, y_n\}$.

5.15.1 $S + (n - S) \frac{p - p\gamma}{1 - \gamma p}$

5.15.2 $N\gamma$

5.15.3 $\rho(N, S) = \left(\frac{\gamma(1-p)}{1-p\gamma} \right)^{\frac{1}{2}}$

5.15.4 (i) $\mathbf{P}(T = k, S = j, N = i) = \binom{j}{k} \tau^k (1-\tau)^{j-k} \binom{i}{j} \gamma^j (1-\gamma)^{i-j} \binom{n}{i} \times p^i (1-p)^{n-i}$ for $k \leq j \leq i \leq n$.

(ii) $\mathbf{P}(N = i | T = k) = \binom{n-k}{n-i} \left(\frac{1-p}{1-p\gamma\tau} \right)^{n-i} \left(\frac{p-p\gamma\tau}{1-p\gamma\tau} \right)^{i-k}$, which is binomial.

5.15.5 $\binom{n-s}{i-s} \left(\frac{p(1-\gamma)}{1-p\gamma} \right)^{i-s} \left(\frac{1-p}{1-p\gamma} \right)^{n-i} \binom{s}{k} \tau^k (1-\tau)^{s-k}$, $k \leq s \leq i$.

5.15.6 Zero

5.17.2 No

5.17.3 $\mathbf{E}(Z) = \frac{1}{2}$; $\text{var}(Z) = \frac{5}{12}$.

5.17.4 $\frac{1}{6}n$; $\frac{5}{36}n$.

5.18.7 Recall the ballot theorem.

5.19.3 $na(b-a)(b-n)/(b^2(b-1))$.

5.19.4 $\mathbf{E}(Z_m) = \frac{b}{a} + \frac{b-1}{a-1} + \dots + \frac{b-m+1}{a-m+1}$.

5.19.5 $\mathbf{P}(X = k | X + Y = j) = \binom{n}{k} \binom{n}{j-k} / \binom{2n}{j}$.

5.20.5 $\mathbf{E}(R_r) = \infty$

5.20.7 For a walk starting at zero, the expected number of visits to zero *including* the first is

$$\frac{1}{|p-q|}. \text{ Hence, for } p < q \text{ and } r > 0, \mathbf{E}(V) = \left(\frac{p}{q} \right)^r \frac{1}{|p-q|}; \text{ for } p < q \text{ and } r < 0, \mathbf{E}(V) = \frac{1}{|p-q|}. \text{ Likewise, if } p > q,$$

$$\mathbf{E}(V) = \begin{cases} \left(\frac{p}{q} \right)^r \frac{1}{|p-q|} & r < 0 \\ \frac{1}{|p-q|} & r > 0. \end{cases}$$

$$\begin{aligned}
 5.20.8 \quad \mathbf{E}(V) &= \sum_n \mathbf{P}(S_X = S_Y = 0) = \sum_n \frac{1}{4^{2n}} \sum_{k=0}^n \frac{(2n)!}{(k!)^2((n-k)!)^2} = \sum_n \frac{1}{4^{2n}} \binom{2n}{n}^2 \\
 &= \sum_n \frac{(2n-1)^2(2n-3)^2 \dots 1^2}{(2n)^2(2n-2)^2 \dots 2^2} \geq \sum_n \frac{1}{2n} = \infty.
 \end{aligned}$$

$$5.21.5 \quad \mathbf{P}(X \geq Y) = \int F_Y(x) f_X(x) dx \geq \int F_X(x) f_X(x) dx = \frac{1}{2}.$$

5.21.7 (i) Let I_j be independent Bernoulli random variables with parameter p . Then

$$\sum_1^m I_j \leq \sum_1^n I_j \text{ for } m \leq n.$$

(ii) Let I_j be independent Bernoulli with parameter p_2 , and let K_j be independent Bernoulli with parameter $\frac{p_1}{p_2}$. Then $I_j K_j$ is Bernoulli with parameter p_1 and $\sum_1^n I_j K_j \leq \sum_1^n I_j$.

Problems

- 1 $f(2, 0) = f(0, 2) = \frac{1}{36}$, $f(1, 0) = f(0, 1) = \frac{8}{36}$, $f(0, 0) = \frac{16}{36}$, $f(1, 1) = \frac{2}{36}$;
 $\text{cov}(X, Y) = -\frac{1}{18}$; $\rho(X, Y) = -\frac{1}{5}$.
- 2 (a) Zero; (b) $\frac{1}{16}$; (c) $\frac{1}{2}$; (d) 1.
- 3 e.g. $X = \pm 1$ with probability $\frac{1}{2}$ each, $Y = |X|$.
- 7 $\text{cov}(U, V) = ac + bd + (ad + bc)$; $\rho(X, Y) = 0$ for many choices of a, b, c, d .
- 8 (i) $\mathbf{P}(\text{correct}) = 2 \left(p - \frac{1}{2} \right)^2 + \frac{1}{2}$; (ii) $\mathbf{P}(\text{correct}) = p^3 + 3p(1-p)^2$.
- 9 (b) $\mathbf{P}(U = m, V = n) = p^{m+1}q^n + q^{m+1}p^n$; $\text{cov}(U, V) = (4pq - 1)/(pq)$;
 $\rho(U, V) = -|p - q|$.
- 10 You need results like $\sum_{i=1}^n i^2 = \frac{1}{3}n(n^2 - 1) + \frac{1}{2}n(n + 1)$ and $\sum_{\substack{i \neq j \\ 1 \leq i, j \leq n}} ij = \left(\sum_{i=1}^n i \right)^2 - \sum_{i=1}^n i^2$.
 Then $\text{cov}(X, Y) = -\frac{n+1}{12}$; $\rho(X, Y) = -\frac{1}{n-1} \rightarrow 0$.
- 11 (b) $a - 4a^2$; (c) $\mathbf{E}(X|Y = 0) = \frac{a}{1-2a}$; $\mathbf{E}(X|Y = 1) = \frac{1}{2}$; (d) $a = \frac{1}{4}$.
- 12 $\text{cov}(U, V) = \left(\frac{5}{6} \right)^2 \left(\frac{7}{6} \right)^2$.
- 13 (a) $\mathbf{E}(|XY|) = \mathbf{E}(|X||Y|) = \mathbf{E}(|X|)\mathbf{E}(|Y|) < \infty$; (b) $\mathbf{E}(X^2)\mathbf{E}(Y^2) < \infty$.
- 14 (a) Yes, when $\theta = 3 - 2\sqrt{2}$; (b) no; (c) yes, when $\theta = \frac{1}{2}(\sqrt{5} - 1)$;
 (d) yes, when $\theta = \frac{1}{2}(3 - 2\sqrt{2})$ and independence holds; (e) yes, when $\frac{\alpha\beta}{1-\beta} = 1$;
 (f) yes, when $\alpha = 6\pi^{-2}$.

$$15 \quad (a) f_X(i) = \begin{cases} \frac{1+\theta}{1-\theta}\theta^{|i|}, & i \neq 0 \\ \frac{2\theta}{1-\theta}, & i = 0 \end{cases}$$

$$(c) f_X(i) = \frac{\theta^{2i+3}}{1-\theta}, \quad i \geq 0 \quad (d) f_X(i) = \frac{\theta^{i+1}}{1-\theta}, \quad i \geq 0$$

$$(e) f_X(i) = \alpha \left(\frac{i\beta}{c-i\beta} - \frac{(i-1)\beta}{c-(i-1)\beta} \right), \quad 1 \leq i \leq c \quad (f) f_Y(j) = \alpha j^{-2}, \quad 1 \leq j.$$

$$17 \quad (c) \mathbf{P}(X_1 = x_1) = \binom{a_1}{x_1} \left(\frac{\sum_2^k a_i}{\sum_2^k x_i} \right) / \left(\frac{\sum_1^k a_i}{\sum_1^k x_i} \right).$$

$$18 \quad (a) \frac{1-p^c}{1-p};$$

$$(b) \mathbf{P}(\min\{X, Y\} > n) = p_1^n p_2^n \text{ so } \mathbf{E}(Z) = \frac{1}{1-p_1 p_2}.$$

$$20 \quad (a) \frac{p\beta}{1-q\beta}; \quad (b) \frac{\alpha p}{\beta q} \frac{\beta^m q^n}{1-\beta^m q^n}.$$

26 Let (x, y, z) take any of the 8 values $(\pm 1, \pm 1, \pm 1)$. Then

$$\begin{aligned} 1 - xy &= |1 - xy| = |(1 - xy)||(-xz)| \quad \text{because } |(-xz)| = 1, \\ &= |(1 - xy)(-xz)| = |yz - xz| \geq \pm(yz - xz). \end{aligned}$$

Now use Corollary 5.3.2 to get the result.

$$27 \quad f_Y(j) = \frac{1}{2n+1}; \quad f_X(i) = \frac{1}{2m+1}.$$

$$28 \quad f(0, 1) = f(1, 2) = f(2, 0) = \frac{1}{3}. \text{ Then } f_X(i) = f_Y(i) = \frac{1}{3}, \text{ and } \mathbf{P}(X < Y) = \frac{2}{3}, \\ \mathbf{P}(Y < X) = \frac{1}{3}.$$

$$29 \quad (a) \mathbf{E}(U_r) = \frac{U^2}{U+V} + \frac{UV}{U+V} \left(1 - \frac{U+V}{UV} \right)^r \rightarrow \frac{U^2}{U+V}.$$

(b) Let T be the number of tosses to the first head,

$$\mathbf{E}(U_{T-1}) = \frac{U^2}{U+V} + \frac{pU^2V^2}{(U+V)(UVp+U+V-p(U+V))}.$$

$$\text{When } U = V = \frac{1}{p}, \text{ this is } \frac{U}{2} \left(1 + \frac{U}{3U-2} \right) \sim \frac{2}{3}U.$$

31 The total $T = \sum_1^k S_i$, where S_i is the score on the i th ball.

$$(a) \mathbf{E}(T) = \frac{1}{2}k(n+1), \quad \mathbf{E}(S_i S_j) = \frac{2}{n(n-1)} \sum_{i>j} ij = \frac{1}{12}(3n+2)(n+1). \text{ Hence,}$$

$$\text{var}(T) = \frac{1}{12}(n+1)k(n-k). \text{ If } M \text{ is the maximum,}$$

$$\mathbf{P}(M = m) = \binom{m-1}{k-1} / \binom{n}{k} \quad \text{without replacement;}$$

$$\mathbf{P}(M = m) = \binom{m}{n}^k - \binom{m-1}{n}^k \quad \text{with replacement.}$$

- 33 (a) $\frac{1}{2}$; (b) $\frac{3\sqrt{5}-1}{6}$; (c) $\frac{5}{6}$.
- 34 Use the argument of Theorem 5.6.7; $pr^3 - r + q = 0$; $r = (-p + (p^2 + 4pq)^{\frac{1}{2}})/(2p)$.
- 35 $\binom{n}{r} \frac{3^{n-r}}{4^n}$; $\sum_{n=r}^{\infty} \binom{n}{r} \frac{3^{n-r} e^{-8} 8^n}{4^n n!} = \frac{e^{-2} 2^r}{r!}$; mean = variance = 12.
- 38 (a) $\mathbf{P}(M \geq r) = \left(\frac{p}{q}\right)^r$; $\mathbf{P}(M = r) = \left(\frac{p}{q}\right)^r - \left(\frac{p}{q}\right)^{r+1}$; $\mathbf{E}(M) = \frac{p}{q-p}$.
- (b) $\mathbf{P}(M = r | S_0 = -k) = \alpha \beta^k \left(1 - \frac{p}{q}\right) \left(\frac{p}{q}\right)^{r+k}$; $\mathbf{P}(M = r) = \frac{\alpha(q-p)}{q-\beta p} \left(\frac{p}{q}\right)^r$;
 $\mathbf{P}(S_0 = -k | M = r) = \left(1 - \frac{\beta p}{q}\right) \left(\frac{\beta p}{q}\right)^k$; $k \geq 0$.
- 40 For a neat method, see Example (6.6.6).
- 41 No

CHAPTER 6

Exercises

- 6.10.7 $\mathbf{E}(s^{T_{a0}}) + \mathbf{E}(s^{T_{aK}})$
- 6.10.8 $(\lambda_2(s))^a$
- 6.10.9 (i) $\mathbf{P}(T_{a0} < \infty) = \begin{cases} 1 & p \leq q \\ \frac{q}{p} & p \geq q \end{cases}$
(ii) $\mathbf{E}(T_{a0} | T_{a0} < \infty) = \frac{a}{|p-q|}$.
- 6.10.10 (i) $\mathbf{E}(s^T) = ps\mathbf{E}(s^{T_{10}}) + qs\mathbf{E}(s^{T_{01}})$;
(ii) $\mathbf{E}(T | T < \infty) = 1 + \frac{1}{|p-q|}$.
- 6.11.1 (0, 2, 2, 4) and (2, 3, 3, 4).
- 6.11.2 (b) $f(x)$ and $g(x)$ have nonnegative coefficients and $\frac{f(x)g(x)}{f(1)g(1)}$
 $= \left(\frac{x(1-x^{12})}{12(1-x)}\right)^2$.
- 6.11.4 Yes, trivially
- 6.11.5 No
- 6.12.10 $n\sigma^2$
- 6.12.13 r is the chance of extinction derived in Example 6.3.16.
- 6.13.7 Use induction.
- 6.13.9 (a) By Jensen's inequality (4.6.14), we have

$$\mathbf{E}(X^2) = \mathbf{E}(X^2 | X > 0)\mathbf{P}(X > 0) \geq (\mathbf{E}(X | X > 0))^2\mathbf{P}(X > 0)$$

$$= \mathbf{E}(X | X > 0)\mathbf{E}(X).$$
- (b) Hence, $\mathbf{E}(Z_n \rho^{-n} | Z_n > 0) \leq \mathbf{E}(Z_n^2 \rho^{-2n})$
- 6.13.10 Let $\mathbf{E}(s^{Z_n^*}) = G_n^*(s)$. Then

$$\mathbf{E}(s^{Z_n^*} t^{Z_{n+m}^*}) = G_{m-1}^*(t)G_n^*(sG_m(t)) \rightarrow sG_{m-1}^*(t)G_m(t)(\rho-1)/(\rho sG_m(t)-1).$$

- 6.14.4** Set $z = y + 1$ in (3) and equate coefficients.
- 6.15.10** For HHH , $\mathbf{E}(X) = 2 + 4 + 8 = 14$; for HTH , $\mathbf{E}(X) = 2 + 8 = 10$; for HHT , $\mathbf{E}(X) = 8$; for THH , $\mathbf{E}(X) = 8$. The others all follow by symmetry from these.
- 6.15.16** $\frac{1}{p} + \frac{1}{p^2q} + \frac{1}{p^3q^2}$ (= 42 in the fair case).
- 6.16.6** Arguing directly

$$\mathbf{E}(T) = \mathbf{E}(X_1 I_a) + \mathbf{E}((X_1 + T') I_a^c) = \mathbf{E}(X_1 I_a) + \mathbf{E}(X_1 I_a^c) + \mathbf{E}(T) \mathbf{E}(I_a^c).$$

Hence,

$$\mathbf{E}(T) = \frac{\mathbf{E}(X_1)}{1 - \mathbf{E}(I_a^c)} = \frac{\mathbf{E}(X_1)}{\mathbf{P}(X_1 \leq a)} = \frac{1}{q(1 - p^a)},$$

if X_1 is geometric.

- 6.16.7** $\mathbf{E}(T) = b + \mathbf{E}(X_1 I_b^c) / \mathbf{P}(X_1 > b) = (1 - p^b) / (qp^b)$, if X_1 is geometric.
- 6.16.8** This is (1) with $a = 0$ and $b = r$.
- 6.16.9** This is (8) with X geometric.
- 6.16.10** $\mathbf{P}(L_n < r) = \mathbf{P}(W > n)$, where W is as defined in (9). Hence,

$$1 + \sum_n s^n \pi_{n,r} = \sum_n s^n \mathbf{P}(W > n) = \frac{1 - \mathbf{E}(s^W)}{1 - s} = \frac{1 - p^r s^r}{1 - s + qp^r s^{r+1}}.$$

Problems

- 1** $\sum \mathbf{P}(X < k) s^k = \sum \mathbf{P}(X \leq k) s^{k+1}$
- 2** (a) $G = \frac{1}{n} \frac{n - s^{n+1}}{1 - s}$ (b) $G = \frac{1}{2n+1} \frac{s^{-n} - s^{n+1}}{1 - s}$; $s \neq 0$
- (c) $G = 1 - (1 - s^{-1}) \log(1 - s)$; $|s| \leq 1$
- (d) $G = 1 - \frac{1}{2} (1 - s) \log(1 - s^{-1}) - \frac{1}{2} (1 - s^{-1}) \log(1 - s)$; $|s| = 1$
- (e) $G = \frac{1 - c}{1 + c} \left\{ 1 + \frac{cs}{1 - cs} + \frac{cs^{-1}}{1 - cs^{-1}} \right\}$; $|s| = 1$.
- 3** (a) A p.g.f. wherever $G_X(s)$ exists; (b) not a p.g.f. (c) a p.g.f. for $|s| < p^{-1}$; (d) a p.g.f. for all s ; (e) a p.g.f. for $|s| \leq 1$; (f) a p.g.f. if $\alpha \log(1 + \beta) = 1$, $\beta < 0$, for $|\beta s| < 0$.
- 5** Let N have p.m.f. $f_N(k) = 2^{-k}$, $k \geq 1$, and $(X_i; i \geq 1)$ be independent and identically distributed with p.g.f. G , then $Y = \sum_1^N X_i$.
- 6** If it were possible, then $1 - s^{11} = (1 - s)R_1(s)R_2(s)$, where R_1 and R_2 are polynomials (with real positive coefficients) of degree five. Because the imaginary roots of unity form conjugate pairs, this is impossible.
- 7** (b) Yes, make it a flat with $f(2) = 0$.
- 9** (b) $G_Y^n(1) + G_Y'(1) - (G_Y'(1))^2 = \text{var}(N)(\mathbf{E}(X))^2 + \mathbf{E}(N)\text{var}(X)$.
- 10** (c) $\frac{1}{3}(G_X(1) + G_X(\omega) + G_X(\omega^2))$, where ω is a complex cube root of unity.

11 $G = G_X$, if and only if $G = \frac{(1 + \mu)^{-1}}{1 - s(1 + \mu)^{-1}}$, so X is geometric.

13 Use conditioning. So $\mathbf{E}(s^N) = \frac{s}{2}\mathbf{E}(s^N) + \frac{s^2}{4}\mathbf{E}(s^N) + \frac{s^2}{4}$;
 $\mathbf{E}(N) = \frac{1}{2}(1 + \mathbf{E}(N)) + \frac{1}{4}(2 + \mathbf{E}(N)) + \frac{1}{2}$; $\mathbf{E}(N) = 6$.

14 (a) $\frac{1}{2}\lambda + 1$. (b) $\mathbf{P}(R = r) = \sum \mathbf{P}(R = r | X = x)\mathbf{P}(X = x) = \sum_{x=r}^{\infty} \frac{cp^x}{x(x+1)}$.

Hence, $(1 - s)G_R(s) = cd(1 - p)s + cs \left(1 - \frac{1}{ps}\right) \log(1 - ps)$,

where $d = \left(\frac{1}{p} \log(1 - p)\right)$. $\mathbf{E}R = \frac{1}{2}cp/(1 - p)$.

15 By the independence $\text{var}(H - T) = \text{var}(H) + \text{var}(T) = \lambda = \text{var}(N)$.

16 With the notation of Problem 15,

$$\mathbf{E}(s^H t^T) = \mathbf{E}(s^H t^{N-H}) = \mathbf{E}((ps + qt)^N) = G_N(ps + qt).$$

If H and T are independent, then

$$G_N(ps + qt) = G_H(s)G_T(t) = G_N(ps + q)G_N(p + qt).$$

Write $s = x + 1$, $t = y + 1$, $G_N(v) = f(v - 1)$ to get $f(px + qy) = f(px)f(qy)$. The only continuous solutions of this are $f(z) = e^{\lambda z}$, so $G_N(s) = e^{\lambda(s-1)}$.

17 $G_{X_n}(s) = \left(\frac{p}{1 - qs}\right)^n = \left(\frac{1 - \frac{\lambda}{n}}{1 - \frac{\lambda s}{n}}\right)^n \rightarrow e^{\lambda(s-1)}$.

18 (b) $\mathbf{E}(s^N) = s + (s - 1)(\exp(se^{-a}) - 1)$; $\mathbf{E}(N) = e^{e^{-a}}$.

19 Do not differentiate $G(s)$!

20 (i) $G_X(s) = \frac{s}{3 - 2s}$. (ii) $G_Y(s) = s \cdot \frac{(m-1)s}{m-s} \cdot \frac{(m-2)s}{m-2s} \cdots \frac{s}{m-(m-1)s}$.

Do not differentiate this to find the mean!

21 Use L'Hopital's rule.

22 $G_n(s) = \frac{s}{2 - G_{n-1}(s)}$; $G_n(1) = 1$; $G'_n(1) = n$.

24 Differentiate.

25 $\left(\frac{a}{1 + a - s}\right)^\lambda$

26 $\mathbf{E}(s^{X+Y} t^{X-Y}) = \exp(\lambda(st - 1) + \mu(st^{-1} - 1))$. Hence, for $X - Y$, $\kappa_r = \lambda + (-)^r \mu$ and for $X + Y$, $\mu^{(k)} = (\lambda + \mu)^k$.

27 $(ps + q)^n (p + qs^{-1})^m s^m = (ps + q)^{m+n}$.

28 (a) $\frac{\alpha p}{1 - [1 - p + p(1 - \alpha)s]t}$; (b) $\frac{p(1 - \alpha)}{1 - \alpha p}$.

29 For all $0 \leq r \leq n$, we have

$$\begin{aligned} \sum_{k=0}^n \binom{n-r}{k} x^k &= (1+x)^{n-r} = \left(1 - \frac{x}{1+x}\right)^r (1+x)^n \\ &= \sum_{k=0}^n \binom{r}{k} \left(\frac{-x}{1+x}\right)^k (1+x)^n. \end{aligned}$$

Because the sums are polynomials in r of degree at most n , it follows that they must be identically equal. Hence, setting $r = -n - 1$ and $x = 1$ gives

$$\sum_{k=0}^n \binom{n+k}{k} 2^{-k+n} = \sum_{k=0}^n \binom{2n+1}{k} = \frac{1}{2} \sum_{k=0}^{2n+1} \binom{2n+1}{k} = 2^{2n}.$$

$$\text{Hence, } \sum_{k=0}^n a_k = \frac{1}{2^{n+1}} \sum_{k=0}^n \binom{n+k}{k} 2^{-k} = \frac{1}{2}.$$

Recognising the p.g.f. of the negative binomial distribution with parameter $\frac{1}{2}$, this says that in a sequence of coin tosses the chance of getting up to n tails before $n+1$ heads equals the chance of getting $n+1$ or more tails before $n+1$ heads equals $\frac{1}{2}$. Now remember the ant of Example 3.7.1.

30 Let $S_n = X_n + Y_n$. Then S_n is a simple random walk with $p = \alpha_1 + \alpha_2$ and $q = \beta_1 + \beta_2 = 1 - p$.

31 S_n is symmetric so (a) $\mathbf{E}(T) = \infty$ and (b) $\mathbf{E}(s^{T_1})|_{s=1} = 1$.

(c) Let $U_n = X_n - Y_n$ and $V_n = X_n + Y_n$, so

$$\begin{aligned} \mathbf{E}(s^{U_1} t^{V_1}) &= \frac{1}{4}(st + st^{-1} + ts^{-1} + s^{-1}t^{-1}) \\ &= \frac{1}{2}(s + s^{-1}) \frac{1}{2}(t + t^{-1}). \end{aligned}$$

Hence, U_n and V_n are independent simple random walks and

$$\begin{aligned} \mathbf{E}(s^{X_r - Y_r}) &= \mathbf{E}(\mathbf{E}(s^{V_r} | T)) \\ &= \mathbf{E}\left(\left(\frac{s + s^{-1}}{2}\right)^T\right) \\ &= \left(F_1\left(\frac{s + s^{-1}}{2}\right)\right)^m, \quad \text{where } F_1(s) = \frac{1 - (1 - s^2)^{\frac{1}{2}}}{s}. \end{aligned}$$

40 $\mathbf{E}(s^{X_m} | X_m > 0) = (\mathbf{E}(s^{X_m}) - \mathbf{P}(X_m = 0))/\mathbf{P}(X_m > 0)$

$$\begin{aligned} &= \left(\left(\frac{p}{1-qs}\right)^m - p^m\right)/(1-p^m) = p^m(e^{-m \log(1-qs)} - 1)/(1 - e^{m \log p}) \\ &= p^m(m \log(1-qs) + O(m^2))/(m \log p + O(m^2)) \\ &\rightarrow \frac{\log(1-qs)}{\log(1-q)} \quad \text{as } m \rightarrow \infty. \quad [O(\cdot) \text{ is defined in Section 7.5}] \end{aligned}$$

41 We know $U(s) = \sum_{k=0}^n u_{2k} s^{2k} = (1 - s^2)^{-\frac{1}{2}}$. Let $r_{2n} = \sum_{k=0}^n u_{2k}$. Then

$$\begin{aligned} \sum s^{2k} r_{2k} &= \frac{U(s)}{1-s^2} = (1-s^2)^{-3/2} = \frac{1}{s} \frac{d}{ds} \left(\frac{1}{(1-s^2)^{\frac{1}{2}}}\right) = \frac{1}{s} \sum \frac{d}{ds} \binom{2k}{k} 4^{-k} s^{2k} \\ &= \sum 2k \binom{2k}{k} 4^{-k} s^{2k-2} = \sum (2k+2) \binom{2k+2}{k+1} 4^{-k-1} s^{2k}. \quad \text{The result follows.} \end{aligned}$$

CHAPTER 7

Exercises

7.11.4 $f(x) = \frac{3}{4} + \frac{1}{4} \cdot 12 \left(x - \frac{1}{2}\right)^2$, so you toss a coin twice and set

$$X = \begin{cases} \left(\frac{U}{4} - \frac{1}{8}\right)^{\frac{1}{3}} + \frac{1}{2} & \text{if you get two heads} \\ U & \text{otherwise.} \end{cases}$$

7.11.5 $\mathbf{P}(Y \leq y) = \mathbf{P}\left(-\log U \leq \left(\frac{y}{\gamma}\right)^\beta\right) = 1 - \exp\left(-\left(\frac{y}{\gamma}\right)^\beta\right)$

7.11.7 (a) $(\pi(1+x^2))^{-1}; -\infty < x < \infty$ (b) $2(\pi(1+x^2))^{-1}; 0 \leq x < \infty$.

7.12.7 First note that the coefficient of x^n in H_n is 1, so $D^n H_n = n!$ Now integrating by parts

$$\int_{-\infty}^{\infty} H_n H_m \phi = [(-)^{m-1} H_n H_{m-1} \phi]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} D H_n (-)^{m-1} D^{m-1} \phi.$$

The first term is zero, and repeated integration by parts gives zero if $m > n$, or

$$\int_{-\infty}^{\infty} \phi D^n H_n = n! \text{ if } m = n.$$

7.12.8 By Taylor's theorem $\phi(x) \sum \frac{t^n H_n(x)}{n!} = \sum t^n (-)^n D^n \phi(x) / n! = \phi(x - t)$.

$$\text{Hence, } \sum \frac{t^n H_n(x)}{n!} = e^{-\frac{1}{2}(x-t)^2 + \frac{1}{2}x^2} = e^{-\frac{1}{2}t^2 + xt}.$$

7.12.9 Set $\phi = -\phi'/x$ in the integral in (6), and integrate by parts again.

7.12.10 $\mathbf{E}(X - t | X > t) = \int_t^{\infty} \frac{1 - F(x)}{1 - F(t)} dx = e^{(\lambda t)^2} \int_t^{\infty} e^{-(\lambda x)^2} dx$
 $= e^{(\lambda t)^2} \pi^{\frac{1}{2}} \lambda^{-1} (1 - \Phi(\lambda + \sqrt{2}))$ and the inequality follows using (3).

7.13.2 Still $\frac{1}{3}$.

7.14.5 The policy is essentially the same with the one difference that $\hat{x} = \hat{y}$.

7.14.6 The new expected cost function λ^* is related to λ by $\lambda^*(x) =$

$$\lambda(x) + m\mathbf{P}(Z > x) = \lambda(x) + m(1 - F(x)). \text{ Then } \frac{\partial \mu}{\partial y} = 0 \text{ yields } 0 = c - h - (h + p + m\lambda) \exp(-\lambda(\hat{y} - a)). \text{ Thence, } \lambda^*(\hat{x}) + c\hat{x} = k + c\hat{y} + \lambda^*(\hat{y}).$$

7.15.6 Let $g(s) = \log s - (s - 2)(s + 1)$. At $s = 1$, we have $g(1) = 2 > 0$; at $s = e^4$, we have $g(e^4) = 4 - (e^4 - 2)(e^4 + 1) = (3 - e^4)(2 + e^4) < 0$. There is thus at least one root. However, $\log s$ lies below its tangent and $s^2 - s - 2$ lies above, so there can be no more than one root in this interval.

7.15.7 Use $\log s \leq s - 1$.

7.16.2 $r(t) = \lambda$. Your part has no memory.

7.16.3 (a) Use Bayes' theorem. (b) $\pi \rightarrow 1$ if $\lambda > \mu$; $\pi \rightarrow 0$ if $\lambda < \mu$; $\pi = p$ if $\lambda = \mu$.

7.16.4 $\frac{d^2}{d\theta^2} \log M(\theta) = \left(\frac{\mathbf{E}(X^2 e^{\theta X}) \mathbf{E}(e^{\theta X}) - (\mathbf{E}(X e^{\theta X}))^2}{(M(\theta))^2} \right).$

$$\text{By Cauchy-Schwarz, } (\mathbf{E}(X e^{\theta X}))^2 = [\mathbf{E}(X e^{\theta X/2} e^{\theta X/2})]^2 \leq \mathbf{E}(X^2 e^{\theta X}) \mathbf{E}(e^{\theta X}).$$

7.16.5 $\mathbf{P}(T > t) = \mathbf{E}(\exp(-\Lambda t)) = M_\Lambda(-t)$. But

$$\frac{dr(t)}{dt} = -\frac{d^2}{dt^2} \mathbf{P}(T > t) = -\frac{d^2}{dt^2} M_\Lambda(-t) < 0, \text{ by (4). Hence, } T \text{ is DFR.}$$

7.16.6 As above $r'(t) = \mathbf{E}(f'_{T_\Lambda})(1 - \mathbf{E}(F_{T_\Lambda})) + (\mathbf{E}(f_{T_\Lambda}))^2$. Now because F_{T_Λ} is DFR, we have $(1 - F_{T_\Lambda})f'_{T_\Lambda} + f_{T_\Lambda}^2 \leq 0$; hence,
 $(\mathbf{E}(f_{T_\Lambda}))^2 \leq (\mathbf{E}(-f'_{T_\Lambda}(1 - F_{T_\Lambda}))^{\frac{1}{2}})^2 \leq \mathbf{E}(-f'_{T_\Lambda})\mathbf{E}(1 - F_{T_\Lambda})$ by Cauchy–Schwarz.
 Hence, $r'(t) \leq 0$.

7.17.4 (i) $\frac{d}{dt} \frac{1}{t} \int_0^t r(v)dv = \frac{r(t)}{t} - \frac{1}{t^2} \int_0^t r(v)dv = \frac{1}{t^2} \int_0^t [r(t) - r(v)]dv > 0$

if $r'(v) > 0$ for all v . Hence, IFR \Rightarrow IFRA. (ii) Use (7.8.6) and Theorem 7.8.7.

7.17.5 (i) $\mathbf{E}(T - t|A_t) = \int_0^\infty \frac{\mathbf{P}(T > t + s)}{\mathbf{P}(T > t)} ds \leq \int_0^\infty \mathbf{P}(T > s) ds$ if NBU.

by Definition 7.7.8 (iii). Hence, NBU \Rightarrow NBUE.

7.17.6 $\mathbf{P}(T > t) = \int_t^\infty \lambda^2 x e^{-\lambda x} dx = (1 + \lambda t)e^{-\lambda t}$. Hence

$$H(t) = -\log(1 + \lambda t) + \lambda t, r(t) = \lambda - \frac{1}{1 + \lambda t}, \text{ and } r'(t) = \frac{1}{(1 + \lambda t)^2} > 0.$$

7.18.3 Uniform on $(0, \frac{1}{2})$.

7.18.4 $\mathbf{E}\{X \wedge (1 - X)\} = \frac{1}{4}$; $\mathbf{E}\{X \vee (1 - X)\} = \frac{3}{4}$.

7.18.5 $\mathbf{E}(\sin \Theta) = 2 \int_0^{\frac{1}{2}} \frac{x}{(x^2 + (1 - x)^2)^{\frac{1}{2}}} dx = \frac{1}{\sqrt{2}}(1 + \log(1 + \sqrt{2})) - 1$.

Likewise, $\mathbf{E}(\cos \Theta) = \frac{1}{\sqrt{2}}(-1 + \log(1 + \sqrt{2})) + 1$, so

$$\frac{\mathbf{E}(\sin \Theta)}{\mathbf{E}(\cos \Theta)} = \frac{\log(1 + \sqrt{2}) + 1 - \sqrt{2}}{\log(1 + \sqrt{2}) - (1 - \sqrt{2})} \simeq 0.36.$$

7.18.7 $\mathbf{E}(\cot \Theta) = \int_0^{\frac{1}{2}} \frac{1}{B(a, b)} x^{a-2}(1 - x)^b dx + \int_{\frac{1}{2}}^1 \frac{1}{B(a, b)} x^a(1 - x)^{b-2} dx$.

7.19.6 Remember (or prove) that $\Gamma(n) = (n - 1)!$ when n is an integer.

7.19.7 By the reflection principle, the number of paths that visit b on the way from $(0, 0)$ to $(2n, 0)$ is the same as the number of paths from $(0, 0)$ to $(2n, 2b)$, namely, $\binom{2n}{n - b}$.

Hence, the probability required is $\binom{2n}{n - b} / \binom{2n}{n}$

$$= \frac{(n!)^2}{(n - b)!(n + b)!} \simeq \frac{n^{2n+1}}{(n - b)^{n-b+\frac{1}{2}}(n + b)^{n+b+\frac{1}{2}}} \text{ using Stirling's formula}$$

$$= \left(1 - \frac{b}{n}\right)^{-n+b-\frac{1}{2}} \left(1 + \frac{b}{n}\right)^{-n-b+\frac{1}{2}} \rightarrow e^{-y^2}.$$

7.19.8 (a) Use Stirling's formula. (b) Take logs. (c) Use the Riemann integral.

7.19.9 The number of paths from $(0, 0)$ to $(2n, 2j)$ is $\binom{2n}{n - j}$; the number from $(0, 0)$ to

$(2r, 0)$ is $\binom{2r}{r}$; and the number from $(2r, 0)$ to $(2n, 2j)$ that do not visit 0 is

$\frac{j}{n - r} \binom{2n - 2r}{n - r + j}$; (recall the reflection principle, or the hitting time theorem). Hence,

the required probability is $\frac{j}{n - r} \binom{2n - 2r}{n - r + j} \binom{2r}{r} / \binom{2n}{n - j} = f_r$. Now use

Stirling's formula and take logs in the usual way.

Problems

- 1 x lies between α and β , $c(\alpha, \beta)^{-1} = \left| \int_{\alpha}^{\beta} (x - \alpha)(\beta - x) dx \right|$.
- 2 $\mathbf{P}(X = x) = \lim_{n \rightarrow \infty} \left(F(x) - F\left(x - \frac{1}{n}\right) \right)$.
- 3 $\text{var}(X) = \frac{B(a+2, b)}{B(a, b)} - \left(\frac{B(a+1, b)}{B(a, b)} \right)^2$.
- 4 $1 = c \int_0^{\pi/2} (\sin x)^{\alpha} (\cos x)^{\beta} dx = \frac{c}{2} B\left(\frac{\alpha+1}{2}, \frac{\beta+1}{2}\right)$.
- 5 $\frac{2}{\pi} \sin^{-1} x^{1/2}$
- 6 $\exp(-\exp(-x))$
- 7 $\mathbf{E}(Y) < \infty$ for $\lambda > 2a > 0$
- 9 $\left(1 - \Phi\left(x + \frac{a}{x}\right)\right) / \left(1 - \Phi(x)\right) = \frac{\phi\left(x + \frac{a}{x}\right)r\left(x + \frac{a}{x}\right)}{\phi(x)r(x)} \rightarrow e^{-a}$,
using the properties of Mills ratio $r(x)$, and $\phi(x)$.
- 11 (i) $F(x) = (b - 4\pi^2 ml_0 x^{-2}) / (b - a)$ for $2\pi \left(\frac{ml_0}{b}\right)^{\frac{1}{2}} \leq x \leq 2\pi \left(\frac{ml_0}{a}\right)^{\frac{1}{2}}$.
(ii) $\mathbf{E}(X) = (4\pi(ml_0)^{\frac{1}{2}}) / (\sqrt{a} + \sqrt{b})$.
- 12 Choose x_l such that $F(x_l) < 2^{-(n+1)}$ and x_u such that $1 - F(x_u) < 2^{-(n+1)}$. Then set $S_n(X) = x_l + r\epsilon$ for $x_l + r\epsilon < X \leq x_l + (r+1)\epsilon$, for all r in $0 \leq r \leq \left\lfloor \frac{x_u - x_l}{\epsilon} \right\rfloor$.
- 14 $\mathbf{P}(X > x) = (1 - 2\sqrt{3}x)^2$, so $f_X(x) = 4\sqrt{3}(1 - 2\sqrt{3}x)$.
- 15 $c = \mathbf{E}(X)$
- 16 $\exp\left(-\frac{2}{3}\lambda t^{\frac{3}{2}}\right)$, $t > 0$
- 17 $F'(x) \propto g(x)$ and $F(\infty) = 1$
- 18 $\kappa_1 = \mu; \kappa_2 = \sigma^2; \kappa_r = 0, r \geq 3$
- 19 $\log M_X(t) = -\log\left(1 - \frac{t}{\lambda}\right) = \sum_{r=1}^{\infty} \frac{t^r}{r\lambda^r}$, so $\kappa_r = \frac{(r-1)!}{\lambda^r}$
- 20 Set $X = \begin{cases} U & \text{if } 0 \leq V \leq \frac{3}{5} \\ U^{\frac{1}{2}} & \text{if } \frac{3}{5} < V \leq \frac{9}{10} \\ U^{\frac{1}{3}} & \text{if } \frac{9}{10} < V \leq 1 \end{cases}$
- 21 (a) Set $u = v/a$ (b) Set $u = b/v$ after differentiating. (c) Integrate $\frac{\partial I}{\partial b} = -2al$ with the boundary condition $I(a, 0) = \pi^{\frac{1}{2}}/(2a)$.
- 22 (a) Set $x = v^2$ to get $M_X(t) = 2al((\gamma - t)^{\frac{1}{2}}, \beta^{\frac{1}{2}}) = \alpha \left(\frac{\pi}{\gamma - t}\right)^{\frac{1}{2}} \exp(-2(\beta(\gamma - t))^{\frac{1}{2}})$, so $\alpha = \left(\frac{\gamma}{\pi}\right)^{\frac{1}{2}} \exp(+2(\beta\gamma)^{\frac{1}{2}})$. (b) $\exp\left(-2\left(-\frac{t}{2}\right)^{\frac{1}{2}}\right)$ for $t \leq 0$.
- 23 Use Problem 22.
- 24 $\mathbf{E}(e^{tX^2}) = \int (2\pi)^{-\frac{1}{2}} \exp\left((2t-1)\frac{x^2}{2}\right) dx = (1-2t)^{-\frac{1}{2}}$. Hence X^2 has the $\chi^2(1)$ density.

25 $\frac{\alpha\mu}{\mu+t} + \frac{\beta\lambda}{\lambda-t}$ for $-\mu < t < \lambda$, where $\alpha + \beta = 1$.

26 $\mathbf{P}(X > k) = \mathbf{P}\left(\frac{\log U}{\log(1-p)} \geq k\right) = \mathbf{P}(U \leq q^k) = q^k$.

27 $X = \begin{cases} U & \text{with probability } \frac{24}{25} \\ \left(\frac{1}{4}U - \frac{1}{8}\right)^{\frac{1}{3}} + \frac{1}{2} & \text{with probability } \frac{1}{25}. \end{cases}$

28 Use Example 7.9.8.

29 $\int_0^\infty f(\theta, x)g(\theta)d\theta = \frac{v}{(v+x)^2}$. This is nonnegative and $\int_0^\infty \frac{v}{(v+x)^2}dx = 1$.

30 $\mathbf{E}(\exp(t(X_n - n)n^{-\frac{1}{2}})) = e^{-tn^{\frac{1}{2}}} \exp(n(e^{-tn^{-\frac{1}{2}}} - 1))$
 $= \exp(-tn^{\frac{1}{2}} + tn^{\frac{1}{2}} + \frac{1}{2}t^2 + O(n^{-\frac{1}{2}})) \rightarrow e^{\frac{1}{2}t^2}$; now use the continuity theorem.

31 $(1-U)^{-d^{-1}}$

32 X_α has mean and variance α , and m.g.f. $\frac{1}{(1-t)^\alpha}$. Hence,

$$M_\alpha(t) = \mathbf{E}\left(\exp\left(\frac{t}{\sqrt{\alpha}}(X_\alpha - \alpha)\right)\right) = e^{-t/\sqrt{\alpha}}\left(1 - \frac{t}{\sqrt{\alpha}}\right)^{-\alpha}.$$

Now for fixed t , and $\left|\frac{t}{\sqrt{\alpha}}\right| < 1$,

$$\begin{aligned} \log M_\alpha(t) &= -\frac{t}{\sqrt{\alpha}} - \alpha \log\left(1 - \frac{t}{\sqrt{\alpha}}\right) \\ &= -\frac{t}{\sqrt{\alpha}} + \alpha \sum_{r=1}^{\infty} \frac{1}{r} \left(\frac{t}{\sqrt{\alpha}}\right)^r \rightarrow \frac{t^2}{2} \text{ as } \alpha \rightarrow \infty. \end{aligned}$$

Using the continuity theorem gives the result.

33 Use Chebyshev's inequality.

CHAPTER 8

Exercises

8.11.3 By definition, $f_{Y|X}(y|x) = f(x, y)/f_X(x) = \frac{1}{(2\pi(1-\rho^2))^{\frac{1}{2}}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right)$.

This is $N(\rho x, 1-\rho^2)$; therefore, $\mathbf{E}(e^{tY}|X) = \exp\left(\rho X t + \frac{1}{2}(1-\rho^2)t^2\right)$.

8.11.4 By conditional expectation,

$$\begin{aligned} \mathbf{E}(e^{sX+tY}) &= \mathbf{E}(\mathbf{E}(e^{sX+tY}|X)) = \mathbf{E}(e^{(s+\rho t)X})e^{\frac{1}{2}(1-\rho^2)t^2} \\ &= \exp\left(\frac{1}{2}\left(s + \rho t\right)^2 + \frac{1}{2}(1-\rho^2)t^2\right) \end{aligned}$$

as above.

8.11.5 $\mathbf{E}(e^{sW+tZ}) = \mathbf{E}\left(\exp\sum(\alpha_i s + \beta_i t)X_i\right) = \exp\left(\frac{1}{2}\sum(\alpha_i s + \beta_i t)^2\right)$. This factorizes as required for the independence if and only if $\sum \alpha_i \beta_i = 0$ (or in geometrical terms, $\alpha \cdot \beta = 0$).

8.11.6 $E(e^{t(aX+bY)}) = \exp\left(\frac{1}{2}t^2(a^2 + 2\rho ab + b^2)\right)$. Hence, $aX + bY$ is $N(0, a^2 + 2\rho ab + b^2)$.

8.12.4 A triangle is feasible if $U < V + W$ and $V < U + W$ and $W < U + V$. In terms of X and Y , this gives (when $X < Y$) the constraints $X < \frac{1}{2}$, $Y - X < \frac{1}{2}$ and $Y > \frac{1}{2}$. A similar possibility arises when $X > Y$. Now a sketch of these two regions shows that they form two triangles with combined area $\frac{1}{4}$, and this is the required probability because (X, Y) is uniform on the unit square.

8.12.6 By symmetry, it is the same as $X_{(1)}$, namely $n(1-x)^{n-1}$; $0 \leq x \leq 1$.

8.12.7 By symmetry, this is the same as the joint density of $X_{(1)}$ and $1 - X_{(n)}$. Now $P(X_{(1)} > x, 1 - X_{(n)} > y) = (1 - x - y)^n$, so $f = n(n-1)(1-x-y)^{n-2}$.

8.12.8 Given neither point is on the diameter, the density of the angle they make at the midpoint of the diameter is given by (1) with $a = \pi$. Hence, the expected area in this case is $\int_0^\pi \frac{2(\pi-x)}{\pi^2} \frac{1}{2} \sin x dx = \frac{1}{\pi}$.

Given one on the diameter and one not, they are jointly uniform on $(0, \pi) \times (-1, 1)$, so the expected area is

$$2 \int_0^\pi \frac{1}{2\pi} \int_0^1 \frac{y \sin x}{2} dx dy = \frac{1}{2\pi}.$$

Hence, the expected area is $\frac{1}{\pi} \left(\frac{\pi}{\pi+2}\right)^2 + \frac{1}{2\pi} \frac{4\pi}{(\pi+2)^2} = \frac{1}{2+\pi}$.

8.13.4 The easy method uses (1) and (2) to see that

$P(A^c \cap B^c) = 1 - P(A \cap B) - P(A \cap B^c) - P(A^c \cap B)$, which gives the answer. The other method observes that Q can be divided into five regions in which (given $C = (x, y)$), $P(A^c \cap B^c)$ takes the values

$$0, 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{y}{l}\right), 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{x}{l}\right), 1 - \frac{2}{\pi} \cos^{-1}\left(\frac{x}{l}\right) - \frac{2}{\pi} \cos^{-1}\left(\frac{y}{l}\right)$$

respectively. Identify the regions and do the integrals.

8.13.5 The easy method allows $b \rightarrow \infty$ in (2). You should also do it via an integral.

8.13.6 Draw a picture with no B -lines to see that the probability of an intersection is

$$\frac{1}{\pi a} \int_0^{\pi/2} \min\{l \cos \theta, a\} d\theta.$$

8.13.7 $\frac{(a-l)(b-l)}{ab}$ for $l \leq a \wedge b$. Evens if $2l^2 - 2(a+b)l + ab = 0$, which implies that the coin has diameter $a + b - (a^2 + b^2)^{\frac{1}{2}}$.

8.14.1 $(2n+3)^{-1}$

8.14.2 $\pi n/(n+1)$

8.14.3 $\pi(n-1)/(n+1)$

8.14.4 Integrate (2) remembering that (1) is a density and $\Gamma(\alpha + \beta)$

$$= \int_0^\infty \lambda^{\alpha+\beta} u^{\alpha+\beta-1} e^{-\lambda u} du.$$

8.14.5 Let X and Y be independent with density $\frac{1}{\Gamma\left(\frac{1}{2}\right)}x^{-\frac{1}{2}}e^{-x}$. Then $U = X + Y$ has density

$$\left(\Gamma\left(\frac{1}{2}\right)\right)^{-2} e^{-u} \int_0^u v^{-\frac{1}{2}}(u-v)^{-\frac{1}{2}} dv = \pi e^{-u} \left(\Gamma\left(\frac{1}{2}\right)\right)^{-2}.$$

The result follows.

8.15.6 Set $(1 + x^2)^{-1} = v$ and use Exercise 5.

8.16.4 $\mathbf{P}(N = n) = \left(1 - \frac{1}{a}\right)^n \frac{1}{a}$ with $\mathbf{E}(N) = a - 1$. Hence, a should be as small as (1) permits.

8.16.5 We choose a so that ae^{-x} is as small as possible (since e^{-X} is uniform). Hence,

$$a = \sup_x \left\{ e^{x - \frac{1}{2}x^2} \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \right\} = \sup_x \left\{ \left(\frac{2}{\pi}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(x-1)^2 + \frac{1}{2}} \right\} = \left(\frac{2e}{\pi}\right)^{\frac{1}{2}}.$$

Thus, we get a variable with density $f_S(x)$ if we set $X = -\log U_1$ whenever $e^{\frac{1}{2}}U_1U_2 < \exp(-(\log U_1)^2/2)$.

8.16.6 Now $-\log U_i$ is exponential with parameter 1, so

$$\mathbf{P}(X \leq x | Y > \frac{1}{2}(X-1)^2) \propto \int_0^x \int_{\frac{1}{2}(v-1)^2}^{\infty} e^{-y} e^{-v} dy dv \propto \int_0^x e^{-\frac{1}{2}v^2} dv.$$

Hence, $f_{X|A} = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} e^{-x^2/2}$.

8.17.3 Use independence of increments.

8.17.4 Conditional on $N(24) = k$, the k calls are independent and uniform over $(0, 24)$. Given $X = x$ and $Y = y$, a call at time U finds you in the shower if $x < U < x + y$, with probability $y/24$. Hence, $\mathbf{P}(\text{a call at } U \text{ finds you in the shower} | N = k) = \mathbf{E}(Y)/24 = p$ (say). Hence, the number Z of calls that finds you in the shower given $N = k$ is binomial with parameters k and p . Hence, $\mathbf{E}(s^Z | N = k) = (ps + 1 - p)^k$; hence, $\mathbf{E}(s^Z) = \mathbf{E}((ps + 1 - p)^N) = \exp(24\lambda(ps + 1 - p)) = \exp(\lambda \mathbf{E}(Y)(s - 1))$.

8.17.5 Argue as in (4). Given $N(t) = n$, then R_1 and R_2 have a trinomial mass function with p.g.f. $\mathbf{E}(x^{R_1}y^{R_2} | N = n) = (p_1x + p_2y + 1 - p_1 - p_2)^n$. Hence, $\mathbf{E}(x^{R_1}y^{R_2})$ factorizes into two Poisson p.g.f.s.

8.17.6 $\lambda \min\{s, t\}$

8.18.5 X_{R-1} is always the smallest of X_1, \dots, X_R . So $\mathbf{P}(X_{R-1} \geq x) = \sum_{r=2}^{\infty} \frac{r-1}{r!} \times (1 - F(x))^r = 1 - Fe^{1-F}$. Hence $\mathbf{P}(X_{R-1} \leq x) = Fe^{1-F}$, with density $f(x) = (1 - F(x)) \exp(1 - F(x)) f_X(x)$.

8.18.6 $T > n$ if and only if $X_1 = X_{(n)}$; by symmetry, therefore, $\mathbf{P}(T > n) = \frac{1}{n}$. Hence,

$$\mathbf{P}(T = n) = \frac{1}{n-1} - \frac{1}{n} = \frac{1}{n(n-1)}, n \geq 2. \text{ When } T = n, X_T = X_{(n)}, \text{ so}$$

$$\mathbf{P}(X_T \leq x) = \sum_{n=2}^{\infty} \frac{(F(x))^n}{(n-1)n}, \text{ as required.}$$

8.18.7 If X_1 represents your loss at some hazard and $(X_r; r \geq 1)$ represents the losses of your successors, then the expected time until someone does worse than you is infinite. The argument is symmetrical, of course, but we do not feel so strongly about our good luck.

Problems

- 1 f is a density if $\frac{1}{2}a + 1 \geq 0$, $\frac{1}{2}b + 1 \geq 0$, $a + b = -\frac{1}{2}$. Independence is impossible.
- 2 $\text{cov}(X, Y) = (1 - ab)/144$.
- 3 (a) $c = 1$; (b) $2e^{-1}$; (c) $\frac{1}{2}$.
- 4 $f_Z(z) = \int_0^z g(z)du = zg(z)$.
- 5 (a) $c = (2\pi)^{-1}$; (b) $\int_0^y (a + y^2)^{-\frac{3}{2}} dy = \frac{y}{a(a + y^2)^{\frac{1}{2}}}$, so X has a Cauchy density.
- 7 $f(x, y) = 4y^3x(1 - x)$ for $0 < x < 1$, $0 < y < \frac{1}{x} \wedge \frac{1}{1 - x}$. Hence, you can simulate X with density $6x(1 - x)$ by forming $U^{\frac{1}{2}}/(U^{\frac{1}{2}} + V^{\frac{1}{2}})$ and accepting it as a value of X if $U^{\frac{1}{2}} + V^{\frac{1}{2}} \leq 1$.
- 9 $6x(1 - x)$
- 10 $\mathbf{E}(e^{sU+tV}) = \mathbf{E}(e^{(s+t)X+(s-t)Y}) = \exp\left(\mu_X(s+t) + \frac{1}{2}\sigma_X^2(s+t)^2 + \mu_Y(s-t) + \frac{1}{2}\sigma_Y^2(s-t)^2\right)$, which factorizes if $\sigma_X = \sigma_Y$.
- 11 Given $Z < 1$, the point $(2U - 1, 2V - 1)$ has the uniform density over the unit disc, namely, $\frac{r}{\pi}$ in polar coordinates. Thus, $Z = R^2$ and $0 < \Theta < 2\pi$. Make the transformation $X = (2 \log R^{-2})^{\frac{1}{2}} \cos \Theta$, $Y = (2 \log R^{-2})^{\frac{1}{2}} \sin \Theta$, with inverse $r^2 = \exp\left(-\frac{1}{2}(x^2 + y^2)\right)$, $\theta = \tan^{-1} \frac{y}{x}$ and $J = \frac{1}{2} \exp\left(-\frac{1}{4}(x^2 + y^2)\right)$.
- The result follows.
- 12 (i) zero for $a < 1$; (ii) $\frac{a\mu}{\lambda + a\mu}$ for $a \geq 1$.
- 13 $ye^{-xy}/(1 - e^{-y^2})$
- 14 Let A be at the top of the melon; let angle AOB be θ , where O is the centre of the melon. Then the probability that all three remain is, when $\frac{\pi}{2} < \theta < \pi$, $(\pi - \theta)/(2\pi)$.
- But $f_\theta(\theta) = \frac{1}{2} \sin \theta$; $0 < \theta < \pi$.
- Hence, \mathbf{P} (all three remain) $= \int_{\pi/2}^{\pi} (\pi - \theta) \frac{\sin \theta}{4\pi} d\theta = \frac{1}{4\pi}$.
- Likewise, \mathbf{P} (any one of the three remains) $= 3\mathbf{P}$ (given one remains)
- $$= 3 \int_{\pi/2}^{\pi} \theta \frac{\sin \theta}{4\pi} d\theta = \frac{3(\pi - 1)}{4\pi}.$$
- 16 $\mathbf{P}(U = X) = \mathbf{P}(Y > X) = \frac{\lambda}{\lambda + \mu}$.
- 17 Use induction
- 18 (a) $\frac{1}{2}(X + Y)$ by symmetry.
- (b) $\mathbf{E}(X|X + Y = V) = \frac{\sigma^2 + \rho\sigma\tau}{\sigma^2 + 2\rho\sigma\tau + \tau^2} V$; $\mathbf{E}(Y|X + Y = V) = \frac{\tau^2 + \rho\sigma\tau}{\sigma^2 + 2\rho\sigma\tau + \tau^2} V$
- 19 $\mathbf{P}(T \geq j + 1) = \mathbf{P}\left(\sum_{i=1}^j X_i < 1\right) = p_j$ (say). Trivially, $p_1 = \frac{1}{1!}$, $p_2 = \frac{1}{2!}$. Now p_j is the volume of the “ j -dimensional pyramid” with apex O and corners $(1, 0, \dots, 0)$, $(0, 1, \dots, 0)$,

etc. Because $p_j = \int_0^1 x^{j-1} p_{j-1} dx$, the result follows by induction. Finally,

$$\mathbf{E}(T) = \sum \mathbf{P}(T \geq j+1) = e.$$

- 20** A triangle is impossible if $X_1 > X_2 + X_3$. This has the same probability as $1 - X_1 > X_2 + X_3$, namely, $\frac{1}{3!}$ by Problem 19. Two more similar constraints give

$$\mathbf{P}(\text{triangle}) = 1 - 3 \cdot \frac{1}{3!} = \frac{1}{2}.$$

- 21** $\mathbf{P}(n(1 - M_n) > x) = \left(\frac{1-x}{n}\right)^n \rightarrow e^{-x}$.

$$\begin{aligned} \mathbf{E}\left(\exp\left(t \sum_1^N X_i\right)\right) &= \mathbf{E}\left(\mathbf{E}\left(\exp\left(t \sum_1^N X_i\right) \mid N\right)\right) \\ &= \mathbf{E}((\mathbf{E}e^{tX_1})^N) \\ &= (1-p) \left(\frac{\mu}{\mu-t}\right) / \left(1 - \frac{p\mu}{\mu-t}\right) \\ &= \frac{(1-p)\mu}{\mu(1-p)-t}. \end{aligned}$$

So Y is exponential with parameter $\mu(1-p)$.

$$\begin{aligned} \mathbf{E}(C(t)X_1) &= \mathbf{E}(X_1 \mathbf{E}(C(t) \mid X_1)) \\ &= \int_t^\infty ut \lambda e^{-\lambda u} du + \int_0^t u \left(\frac{1}{\lambda} - \frac{1}{\lambda} e^{-\lambda(t-u)}\right) \lambda e^{-\lambda u} du \\ &= t^2 e^{-\lambda t} + \frac{t e^{-\lambda t}}{\lambda} + \frac{1}{\lambda^2} - \frac{t e^{-\lambda t}}{\lambda} - \frac{e^{-\lambda t}}{\lambda^2} - \frac{1}{2} t^2 e^{-\lambda t}. \end{aligned}$$

So $\text{cov}(C(t), X_1) = \frac{1}{2} t^2 e^{-\lambda t}$.

- 24** Let the condition be \bar{A} . Now we notice that

$$\mathbf{P}(X \leq x, A) \propto \int^x x^{\alpha-1} e^{-\frac{\alpha-1}{\alpha}x} e^{-\frac{x}{\alpha}} dx = x^{\alpha-1} e^{-x}.$$

Hence, the result is true by the rejection method Example 8.16, provided that

$\left(\frac{ex}{\alpha}\right)^{\alpha-1} e^{-\frac{\alpha-1}{\alpha}x} \leq 1$ for $x \geq 0$. Because $\alpha - 1 \geq 0$, this is equivalent to

$\log\left(\frac{x}{\alpha}\right) \leq \frac{x}{\alpha} - 1$, so the result holds. This clearly provides a method for simulating gamma random variables, given a supply of uniform random variables.

- 25** Hint: (X_1, \dots, X_n) has the same joint distributions as, (X_n, \dots, X_1) .

- 26** (a) Recalling Theorem 8.4.6 on quotients gives the required density as

$$\int_{-\infty}^{\infty} \frac{1}{|u|} \frac{u^2}{w^2} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(u^2 + \frac{u^2}{w^2}\right)\right) du = 2 \cdot \frac{1}{2\pi w^2} \cdot \frac{1}{(1+1/w^2)}, \text{ as required.}$$

$$\begin{aligned} \text{(b) } \mathbf{E}(e^{tX_1 X_2}) &= \mathbf{E}(\mathbf{E}(e^{tX_1 X_2} \mid X_1)) = \mathbf{E}(e^{\frac{1}{2}t^2 X_2^2}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2(1-t^2)} dx \\ &= (1-t^2)^{-\frac{1}{2}}. \end{aligned}$$

$$\text{Hence, } \mathbf{E}e^{t(X_1 X_2 + X_3 X_4)} = \frac{1}{1-t^2} = \frac{1}{2} \left[\frac{1}{1-t} + \frac{1}{1+t} \right] = \frac{1}{2} (\mathbf{E}(e^{tY}) + \mathbf{E}(e^{-tY})),$$

where Y is exponential with parameter 1. The result follows.

28 Let $V = \max \{U_1, \dots, U_Y\}$. First, we notice that by conditioning on Y

$$F_V = \mathbf{P}(\max \{U_1, \dots, U_Y\} \leq v) = \sum_{y=1}^{\infty} \frac{v^y}{(e-1)y!} = \frac{e^v - 1}{e-1}.$$

Now let us find the m.g.f. of Z ,

$$\mathbf{E}(e^{tZ}) = \mathbf{E}(e^{tX})\mathbf{E}(e^{-tV}) = \frac{(e-1)e^{t-1}}{1-e^{t-1}} \int_0^1 e^{-tv} \frac{e^v}{e-1} dv = \frac{1}{1-t}.$$

Hence, Z is exponential with parameter 1. Alternatively, you can find F_V directly.

29 $\frac{X}{Z}$ has a beta density.

30 Consider their joint m.g.f. $\mathbf{E}(e^{sX+tZ}) = \mathbf{E}(e^{(s+t\rho)X+t(1-\rho^2)^{\frac{1}{2}}Z}) = e^{\frac{1}{2}(s^2+2\rho st+t^2)}$, as required.

35 $f(x) = e^{-x}$. Calculate:

$$\begin{aligned} \mathbf{E}(e^{tU(X_1+X_2)}) &= \mathbf{E}\left(\mathbf{E}(e^{tU(X_1+X_2)}|U)\mathbf{E}\left(\frac{1}{(1-Ut)^2}\right)\right) = \int_0^1 \frac{1}{(1-ut)^2} du \\ &= \frac{1}{1-t} = \mathbf{E}(e^{tY}). \end{aligned}$$

CHAPTER 9

Exercises

9.11.1 (a) Column sums are one, as well as row sums. Hence, $\pi_i = \frac{1}{8}$ satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, and so $\mu_0 = 8 = \mu_V$. (b) $\mathbf{E}(X) = 1$ by the same argument. (c) $\mathbf{E}(T)$ is different. We use the following device. T is the sum of the M steps at which the walk moves to a different vertex and the steps at which it does not move. The number N of nonmoving steps before leaving O has the same expectation (and distribution) as the number at every other vertex on the way from O to V , so $\mathbf{E}(T) = \mathbf{E}(M)\mathbf{E}(N)$. By the example, $\mathbf{E}(M) = 1 + (\alpha + \beta + \gamma) \times (\alpha^{-1} + \beta^{-1} + \gamma^{-1})$, and it is easy to find that $\mathbf{E}(N) = \frac{\delta}{1-\delta}$. Hence $\mathbf{E}(T) = \delta((\alpha + \beta + \gamma)^{-1} + \alpha^{-1} + \beta^{-1} + \gamma^{-1})$.

9.11.2 Consider a random walk on a unit square that takes x -steps with probability p and y -steps with probability q . Then if T is the first passage time from $(0, 0)$ to $(1, 1)$, arguments similar to those of the example show that

$$\mathbf{E}(s^T) = \frac{U_V(s)}{U(s)},$$

where

$$U(s) = \frac{1}{2} \left(\frac{1}{1-s^2} + \frac{1}{1-(p-q)^2s^2} \right)$$

and

$$U_V(s) = \frac{s^2}{2} \left(\frac{1}{1-s^2} - \frac{(p-q)^2}{1-(p-q)^2s^2} \right),$$

which yields $\mathbf{E}(T)$ after some plod. More simply, by conditional expectation we have $\mathbf{E}(T) = 1 + p(1 + p\mathbf{E}(T)) + q(1 + q\mathbf{E}(T))$, which yields $\mathbf{E}(T) = p^{-1} + q^{-1}$. If the walk can wait at vertices with probability r , then by the same device as used in (1),

we find $\mathbf{E}(T) = \frac{r}{p} + \frac{r}{q}$. Now we recognise that the question is equivalent to this problem with $p = \alpha, q = \beta, r = \gamma$.

$$\begin{aligned} \mathbf{9.12.6} \quad LHS &= \mathbf{P}(Y_{n_r} = k, Y_{n_{r-1}} = k_1, \dots, Y_{n_1} = k_{r-1}) / \mathbf{P}(Y_{n_{r-1}} = k_1, \dots, Y_{n_1} = k_{r-1}) \\ &= \frac{\mathbf{P}(X_{-n_r} = k, \dots, X_{-n_1} = k_{r-1})}{\mathbf{P}(X_{-n_{r-1}} = k_1, \dots, X_{-n_1} = k_{r-1})} \\ &= \frac{\mathbf{P}(X_{-n_r} = k | X_{-n_{r-1}}) \mathbf{P}(X_{-n_1} = k_{r-1}, \dots | X_{-n_{r-1}})}{\mathbf{P}(X_{-n_1} = k_{r-1}, \dots | X_{-n_{r-1}})} \\ &= \mathbf{P}(Y_{n_r} = k | Y_{n_{r-1}}), \end{aligned}$$

where we used (9.1.8) at the crucial step. In equilibrium,

$$\begin{aligned} q_{ij} &= \mathbf{P}(Y_2 = j, Y_1 = i) / \mathbf{P}(Y_1 = i) \\ &= \mathbf{P}(X_{-1} = i | X_{-2} = j) \mathbf{P}(X_{-2} = j) / \mathbf{P}(Y_1 = i) = p_{ji} \pi_j \pi_i^{-1}. \end{aligned}$$

9.13.4 No for (a) because of periodicity. Yes for (b).

$$\mathbf{9.13.5} \quad (\text{a}) \mathbf{E}(X_{n+1} | X_n) = 1 - \frac{2}{m} X_n + 1$$

$$(\text{b}) \mathbf{E}(X_n) = \left(1 - \frac{\alpha + \beta}{m}\right)^n \left(\mathbf{E}(X_0) - \frac{\beta m}{\alpha + \beta}\right) + \frac{\beta m}{\alpha + \beta}.$$

9.14.8 Because v_n is a renewal sequence, there is a Markov chain V_n such that $v_n = \mathbf{P}(V_n = 0 | V_0 = 0)$. Let U_n and V_n be independent. Then $((U_n, V_n); n \geq 0)$ is a Markov chain and $u_n v_n = \mathbf{P}((U_n, V_n) = (0, 0) | (U_0, V_0) = (0, 0))$, thus $(u_n v_n; n \geq 0)$ is a renewal sequence.

9.14.9 Consider the chain $(U_{nd}; n \geq 0)$.

9.14.10 If $B_n > 0$, then $B_n - 1 = B_{n+1}$, and if $B_n = 0$, then B_{n+1} is the time to the next event, less the elapsed unit of time. Hence, B is a Markov chain with

$$p_{i, i-1} = 1; \quad i > 0$$

and

$$p_{0j} = f_X(j+1) = \mathbf{P}(X = j+1).$$

Hence, for a stationary distribution π with $\pi(s) = \sum_i s^i \pi_i$,

$$\pi_j = \pi_{j+1} + \pi_0 f_X(j+1),$$

whence

$$\pi(s) = \pi_0 \frac{G_X(s) - 1}{s - 1},$$

and so if $\sum_i \pi_i = 1$,

$$\pi(s) = \frac{1}{\mathbf{E}(X)} \frac{1 - G_X(s)}{1 - s} \quad \text{if } \mathbf{E}(X) < \infty.$$

Hence,

$$\pi_i = \frac{\mathbf{P}(X > i)}{\mathbf{E}(X)}.$$

9.14.11 The transition probabilities of U reversed are

$$q_{i,i-1} = \frac{1 - F_i}{1 - F_{i-1}} \cdot \frac{\pi_{i-1}}{\pi_i} = 1; \quad i > 0$$

and those of B reversed are

$$q_{i,i+1} = 1 \cdot \frac{\pi_{i+1}}{\pi_i} = \frac{1 - F_{i+1}}{1 - F_i}; \quad i \geq 0.$$

Hence, U reversed is B and B reversed is U .

9.15.1 Using (b) and (c) shows that j is persistent.

9.15.2 Follows from (b).

9.15.3 By assumption, $p_{ij}(n) > 0$ and $p_{ji}(m) > 0$ for some finite n and m . Hence, $p_{jj}(m+r+n) \geq p_{ji}(m)p_{ii}(r)p_{ij}(n)$.

Now sum over r to get $\sum p_{jj} = \infty$ if $\sum p_{ii}(r) = \infty$. So if i is persistent so is j . Interchange the roles of i and j .

If j has period t , let $r = 0$ to find that when $p_{jj}(m+n) \geq 0$, $m+n$ is a multiple of t . Hence, the right-hand side is nonzero only when r is a multiple of t , so i has period t .

9.16.8 With $HTH = 1$, $HHH = 2$, we have

$$p_{12}(1) = 0, p_{12}(2) = p^2 = \frac{1}{4} = p_{22}(2) = p_{21}(2) = p_{11}(2)$$

$$p_{22}(1) = 0 = p_{21}(1) = p_{11}(1); p_{22}(1) = p = \frac{1}{2}.$$

Hence,

$$\mu_{12} = \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{4}\right) 8 = 12$$

and

$$\mu_{21} = (1 + p^2 - p^2) 8 = 8.$$

9.16.9 We set $HHH = 1$, $HTH = 2$, then

$$\phi_{s1} = \frac{10 + 12 - 14}{8 + 12} = \frac{3}{10}.$$

9.16.10 Here $TTH = 1$, $HHH = 2$. Calculate $\mu_{s1} = 8$. Now

$$p_{11}(1) = p_{11}(2) = p_{21}(1) = p_{21}(2) = p_{12}(1) = 0,$$

$$p_{12}(2) = \frac{1}{4}, p_{22}(1) = \frac{1}{2}, p_{22}(2) = \frac{1}{4}.$$

Hence,

$$\phi_{s1} = \frac{14 + 8 - 8}{12 + 8} = \frac{7}{10}.$$

9.17.4 $\mathbf{P}(T_1 < \infty) = \mathbf{P}(X(t) > 0) = 1 - \exp(-\Lambda(t))$. Now let $t \rightarrow \infty$.

9.17.5 $\mathbf{E}(s^{X(t)}) = \exp\left((z-1) \int_0^t \lambda e^{-\lambda u} du\right) \rightarrow e^{z-1} = e^{-1} \sum z^k (k!)^{-1}$, and we use the continuity theorem for p.g.f.s.

9.17.6 By conditional expectation

$$\mathbf{E}(e^{\theta Z}) = \mathbf{E}(\mathbf{E}(e^{\theta Z} | X)) = \mathbf{E}((\mathbf{E}e^{\theta Y})^X).$$

You can also get this using forward equations with a lot more work.

$$9.17.7 \quad \mathbf{E}(z^X) = \mathbf{E}(\mathbf{E}(z^X|Y)) = \mathbf{E}\left(\exp\left[(z-1)\int_0^t Y(u)du\right]\right)$$

$$G'_X(z)|_{z=1} = M'(0) = \int_0^t \mathbf{E}(Y(u))du.$$

$$\begin{aligned} G''_X(1) + G'_X(1) - (G'_X(1))^2 &= M''(0) - (M'(0))^2 + M'(0) \\ &= \text{var} \int_0^t Y(u)du + \int_0^t \mathbf{E}(Y(u))du. \end{aligned}$$

$$9.18.2 \quad \exp\left(\frac{\lambda}{\mu}(z-1)(1-e^{-\mu t})\right) \rightarrow \exp\left(\frac{\lambda}{\mu}(z-1)\right)$$

9.18.3 The forward equations are:

$$p'_n(t) = \lambda p_{n-1}(t) + \mu(n+1)p_{n+1}(t) - (\lambda + \mu n)p_n(t).$$

$$\text{Set } G = \sum s^n p_n \text{ to get } \frac{\partial G}{\partial t} = (s-1)\left(\lambda G - \mu \frac{\partial G}{\partial s}\right).$$

$$9.18.4 \quad \pi_k = \frac{1}{2(k!)} \left(1 + \frac{k\mu}{\lambda}\right) \left(\frac{\lambda}{\mu}\right)^k \exp\left(-\frac{\lambda}{\mu}\right).$$

$$9.19.10 \quad \mathbf{E}(s^{X(t)}) = e^{-\delta t} + \int_0^t \delta e^{-\delta x} \exp\left(\frac{\nu}{\mu}(s-1)(1-e^{-\mu x})\right) dx,$$

where ν is the immigration rate, μ the death rate and δ the disaster rate. Hence,

$$\lim_{t \rightarrow \infty} \mathbf{E}(s^{X(t)}) = \int_0^1 \frac{\delta}{\mu} \exp\left(\frac{\nu}{\mu}(s-1)(1-y)\right) y^{(\delta/\mu)-1} dy,$$

where we have set $e^{-\mu x} = y$ in the integrand. Differentiating with respect to s and setting $s = 1$ gives the stationary mean $\nu/(\delta + \mu)$.

9.20.9 The forward equations are

$$\frac{dp_n(t)}{dt} = \lambda_{n-1} p_{n-1}(t) - \lambda_n p_n(t).$$

From (3), we have $f_{n+1}(t) = \sum_{k=1}^{\infty} p'_k(t) = \lambda_n p_n(t)$, so

$$M_{n+1}(\theta) = \lambda_n \int_0^{\infty} p_n(t) e^{\theta t} dt, \text{ as required.}$$

9.20.10 Use (5) and partial fractions.

$$9.21.3 \quad \eta = \lim_{t \rightarrow \infty} \mathbf{P}(X(t) = 0) = \lim_{t \rightarrow \infty} G(0, t)$$

9.21.4 $\mathbf{P}(T > t) = 1 - G(0, t)$. If $\lambda < \mu$ then extinction is certain, so

$$\mathbf{E}(T) = \int_0^{\infty} \mathbf{P}(T > t) dt = \int_0^{\infty} \frac{(\mu - \lambda) \exp((\lambda - \mu)t)}{\mu - \lambda \exp((\lambda - \mu)t)} dt = \frac{1}{\lambda} \log \frac{\mu}{\mu - \lambda}.$$

However, if $\lambda > \mu$, then $\mathbf{P}(T < \infty) = \eta = \frac{\mu}{\lambda}$, so

$$\begin{aligned} \mathbf{E}(T|T < \infty) &= \int_0^{\infty} \left(1 - \frac{\mathbf{P}(T \leq t)}{\eta}\right) dt = \int_0^{\infty} 1 - \frac{G(0, t)}{\eta} dt \\ &= \int_0^{\infty} \frac{(\lambda - \mu) \exp((\mu - \lambda)t)}{\lambda - \mu \exp((\mu - \lambda)t)} dt = \frac{1}{\mu} \log \frac{\lambda}{\lambda - \mu}. \end{aligned}$$

9.21.5 Condition on events in $(0, h)$ (and use the fact that if the first individual splits the replacements act independently) to get

$$\eta(t + h) = 1 \cdot \mu h + \eta(t)(1 - \lambda h - \mu h) + (\eta(t))^2 \lambda h + o(h).$$

Rewrite the equation as $\frac{d\eta}{(1 - \eta)(\mu - \lambda\eta)} = dt$ and integrate to get $\eta(t) = G(0, t)$, as expected. $\mathbf{P}(X(t) = 0 | X(s) = 0) = \frac{\eta(t)}{\eta(s)}$.

9.21.6 $\mathbf{E}(s^{X(t)} | X(t) > 0) = \frac{G(s, t) - G(0, t)}{1 - G(0, t)} \rightarrow \frac{(\mu - \lambda)s}{\mu - \lambda s}$.

9.21.7

$$p_{i, i+1}(h) = (v + \lambda i)h, \quad p_{i, i-1}(h) = \mu i h$$

$$p_{ii}(h) = 1 - (v + \lambda i)h - \mu i h + o(h).$$

Hence,

$$\frac{dp_n}{dt} = \mu(n + 1)p_{n+1} - (v + (\lambda + \mu)n)p_n + (v + \lambda(n - 1))p_{n-1}.$$

Setting $\frac{dp_n}{dt} = 0$, you can check that the given π_k satisfies the resulting equations.

Problems

- 1** (α) X_n is, with $p_{ij} = \mathbf{P}(X_1 = j)$. (β) S_n is, with $p_{ij} = \mathbf{P}(X_1 = j - i)$. (γ) M_n is, with $p_{ij} = \mathbf{P}(X_1 = j)$; $j > i$, $p_{ii} = \mathbf{P}(X_1 \leq i)$. (δ) L_n is, with $p_{ij} = \mathbf{P}(X_1 = j)$, $j < i$, $p_{ii} = \mathbf{P}(X_1 \geq i)$. (ϵ) K_n is not a Markov chain.
- 2** X is persistent; S is transient unless $\mathbf{P}(X_1 = 0) = 1$; M is absorbing if X_1 is bounded, transient otherwise; L is absorbing.
- 3** Only if the period is 2.
- 4** (a) Not necessarily if X and Y are periodic.
- 5** Check that $\mathbf{P}(X_{k+1} = j | X_0, X_1, \dots, X_k, X_n) = \mathbf{P}(X_{k+1} = j | X_k, X_n)$ by expanding the conditional probabilities.
- 6** These all work by expanding the appropriate conditional probabilities and rearranging them.
- 7** It is Markov because coins are independent. Then $p_{i, i+1} = p$; $0 \leq i \leq 8$, $p_{90} = p$, $p_{ii} = q$; $\pi_i = 10^{-1}$.
- 10** $\rho^n u_n = \sum_1^n \rho^{n-k} u_{n-k} \rho^k f_k$, so v_n satisfies $v_n = \sum_1^n f_k^* v_{n-k}$ with $f_k^* = \rho^k f_k$. Hence, it is a renewal sequence provided $\sum \rho^n f_n = 1$. It follows that there is a Markov chain such that $v_n = \mathbf{P}(X_n = s) \rightarrow \pi_s$ as $n \rightarrow \infty$. Hence, $\rho^n u_n \rightarrow \pi_s = c$.
- 11** $Z_n = (X_n, X_{n+1}, \dots, X_{n+m-1})$ is also a persistent chain.
- 13** The Markov property is preserved at each first passage time.
- 14** In the obvious notation, we require $Q^2 = \frac{1}{4}(P_X + P_Y)^2 = Q(2)$; multiplying out $(P_X + P_Y)^2$ shows that this requires $(P_X - P_Y)^2 = 0$. Hence, W is Markov if $P_X = P_Y$.
- 17** No. Pick $j \neq s \neq i$, then $\mathbf{P}(X_{n+1} = j | X_n = s, X_{n-1} = i) = p_{ij}(2) \neq \mathbf{P}(X_{n+1} = j | X_n = s)$.
- 18** The lack-of-memory of the geometric distribution means that X is a Markov chain. If $X_n = i$ then X_{n+1} is the survivors S_n of X_n plus the new arrivals. The probability that a geometric lifetime survives one step is $p = 1 - q$, so S_n is binomial with parameters X_n and p .

Hence $p_{ij} = \mathbf{P}(S_n + Y_n = j)$. In equilibrium X_{n+1} and X_n have the same distribution, so

$$\begin{aligned}\mathbf{E}(s^{X_{n+1}}) &= G(s) = \mathbf{E}(s^{S_n + Y_n}) \\ &= \mathbf{E}(s^{Y_n})\mathbf{E}(s^{S_n} | X_n) = \mathbf{E}(s^{Y_n})\mathbf{E}((ps + 1 - p)^{X_n}) \\ &= e^{\lambda(s-1)}G(ps + 1 - p) = e^{\lambda/q(s-1)}\end{aligned}$$

after a simple induction. So the stationary distribution is Poisson with parameter λ/q , just *after* the fresh particles. The stationary distribution just *before* the fresh particles is Poisson with parameter $\lambda p/q$.

19 We seek a stationary distribution that must satisfy

$$\pi_j = \pi_0 f_j + \sum_{i=1}^{j+1} \pi_i f_{j-i+1} \text{ for } j \geq 0.$$

The sum on the right is nearly a convolution, so we introduce $\pi(s) = \sum \pi_i s^i$ and $G(s) = \sum f_i s^i$ to get

$$\pi(s) = \pi_0 G(s) + \frac{1}{s}(\pi(s) - \pi_0)G(s).$$

Hence,

$$\pi(s) = \frac{\pi_0(s-1)G(s)}{s-G(s)},$$

which has $\pi(1) = 1$ if $\pi_0 > 0$ and $G'(1) < 1$.

20 Seek a stationary distribution that satisfies

$$\pi_k = \sum_{j=k-1}^{\infty} \pi_j f_{j-k+1} \quad \text{for } k \geq 1, \text{ with } \pi_0 = \sum_{j=0}^{\infty} \pi_j \sum_{i=j+1}^{\infty} f_i.$$

Let $\sum \theta^r f_r = G(\theta)$. In an optimistic spirit, we seek a solution of the form $\pi_j = (1 - \theta)\theta^j$, giving

$$\theta^k = \sum_{j=k-1}^{\infty} \theta^j f_{j-k+1} = \sum_{r=0}^{\infty} \theta^{k+r-1} f_r = \theta^{k-1} G(\theta),$$

and

$$1 = \sum_{j=0}^{\infty} \theta^j \sum_{i=j+1}^{\infty} f_i = \frac{1 - G(\theta)}{1 - \theta}.$$

These both reduce to $G(\theta) = \theta$. Hence, π_j exists if $G(\theta) = \theta$ has a root less than 1. By convexity, it does if $G'(1) > 1$, as required.

22 Recalling Example 9.3.17, we define the event $A(t)$ as we did there; that is, as the event that the chain follows a path consistent with $T = t$. Then the condition of the problem can be rewritten as

$$\mathbf{P}(T = t, A(t), X_t = j) = \text{zero or } \mathbf{P}(X_t = j, A(t)).$$

Now the proof follows exactly the same route as Example 9.3.17.

23 π must satisfy $\pi_j = p\pi_{j-2} + q\pi_{j+1}$ with $\pi_0 = q\pi_0 + q\pi_1$ and $\pi_1 = q\pi_2$. Hence,

$s\pi(s) = ps^3\pi(s) + q(\pi(s) - \pi_0) + qs\pi_0$. This gives $\pi(s) = \frac{q(s-1)\pi_0}{s-q-ps^3}$; now insist that $\pi_j \geq 0$ and $\pi(1) = 1$.

24 (a) $\mathbf{P}(X(0) = 1 | X(-\tau) = i, X(\tau) = j)$

$$\begin{aligned} &= \mathbf{P}(X(-\tau) = i, X(0) = 1, X(\tau) = j) / \mathbf{P}(X(-\tau) = i, X(\tau) = j) \\ &= \frac{\mathbf{P}(X(-\tau) = i) p_{i1}(\tau) p_{1j}(\tau)}{\mathbf{P}(X(-\tau) = i) p_{ij}(2\tau)} \rightarrow \frac{\pi_1 \pi_j}{\pi_j}. \end{aligned}$$

(b) $\mathbf{P}(X(s+t) = X(s) = 1) - \mathbf{P}(X(s) = 1)\mathbf{P}(X(s+t) = 1)$ (c) $\rightarrow \pi_1(p_{11}(t) - \pi_1)$, as $s \rightarrow \infty$.

25 (a) Yes, with parameter $\lambda + \mu$; (b) No.

26 $\Lambda(s)$

28 Let $A(t)$ be any event defined by $X(s)$ for $s \leq t$. Then by the Markov property

$$\mathbf{P}(Y_{n+1} = k | Y_n = j, T_n = t, A(T_n)) = \mathbf{P}(Y_{n+1} = k | Y_n = j, T_n = t).$$

Hence,

$$\begin{aligned} \mathbf{P}(Y_{n+1} = j | Y_n = i, A(T_n)) &= \int_0^\infty \mathbf{P}(Y_{n+1} = j | Y_n = i, T_n = t) f_{T_n}(t) dt \\ &= \mathbf{P}(Y_{n+1} = j | Y_n = i), \quad \text{so } Y \text{ is Markov} \\ &= \int_0^\infty p_{ij}(t) \lambda e^{-\lambda t} dt = q_{ij}, \quad \text{say.} \end{aligned}$$

Then $\sum \pi_i q_{ij} = \int \sum \pi_i p_{ij}(t) \lambda e^{-\lambda t} dt = \int \pi_j \lambda e^{-\lambda t} dt = \pi_j$, as required.

$$29 \quad \mathbf{E}(X(t)) = I e^{(\lambda - \mu)t} \quad \text{var}(X(t)) = \begin{cases} 2I\lambda t; & \lambda = \mu \\ I \frac{\lambda + \mu}{\lambda - \mu} e^{(\lambda - \mu)t} (e^{(\lambda - \mu)t} - 1); & \lambda \neq \mu. \end{cases}$$

Further Reading

There are many attractive books on probability. To compile a list as short as this requires regrettably ruthless selection from their number.

Intermediate Probability

If you want to read further at an intermediate level, then high on your list should be the classic text: Feller, W. (1968) *An introduction to probability theory and its applications*, Vol. I (3rd edn.) John Wiley, New York.

Other books at this level or a little above include:

Grimmett, G.R. and Stirzaker, D.R. (2001) *Probability and random processes* (3rd edn.) Clarendon Press, Oxford.

(2001) *One thousand exercises in probability*, Clarendon Press, Oxford.

Ross, S.M. (2003) *Introduction to probability models* (8th edn.) Academic Press, Orlando.

Combinatorics

A classic text on combinatorics for probabilists is:

Whitworth, W.A. (1901) *Choice and Chance* (5th edn.) reprinted 1948 by Hafner, New York.

Recent introductions include:

Anderson, I. (1974) *A first course in combinatorial mathematics*, Clarendon Press, Oxford.

Hall, M. (1967) *Combinatorial theory*, Blaisdell, Waltham, Mass.

Slomson, A. (1991) *An introduction to combinatorics*, Chapman and Hall, London.

Advanced Probability

To advance in probability requires the student to plunge into measure theory. Excellent texts at this level include:

Billingsley, P. (1995) *Probability and measure* (3rd edn.) John Wiley, New York.

Dudley, R.M. (1989) *Real analysis and probability*, Wadsworth, Belmont, Calif.

Durrett, R. (1996) *Probability: theory and examples*, (2nd edn.) Duxbury, Belmont, Calif.

Feller, W. (1971) *An introduction to probability theory and its applications*, Vol II (2nd edn.) John Wiley, New York.

Laha, R.G. and Rohatgi, V.K. (1979) *Probability theory*, John Wiley, New York.

Shiryayev, A.N. (1984) *Probability*, Springer, New York.

Williams, D (1991) *Probability with martingales*, Cambridge University Press.

Markov Chains and Other Random Processes

Most of the above books contain much material on Markov chains and other random processes at their own levels. However, mention should be made of the classic text:

Doob, J.L. (1953) *Stochastic processes*, John Wiley, New York.

History

Finally, if you wish to find out more about the origins of probability read:

Hald, A. (1990) *A history of probability and statistics and their applications before 1750*, John Wiley, New York.

Index of Notation

$B(n, p)$	binomial distribution
$\text{cov}(X, Y)$	covariance
$f(x), f_X(x)$	mass or density function
$f_{X Y}(x y)$	conditional mass or density
$f_{X,Y}(x, y)$	joint mass or density
i	$\sqrt{-1}$
i, j, k, m, n	indices
$\max(x, y) = x \vee y$	maximum of x and y
$\min(x, y) = x \wedge y$	minimum of x and y
$p, p_i, p_{ij}, p(t), p_{ij}(t)$	probabilities
x^+	$\max(x, 0)$
$\text{var}(X)$	variance
$ A $	cardinality or size of A
A^c	complement of A
$B(a, b)$	beta function
$F(x), F_X(x)$	distribution function
$F_{X Y}(x y)$	conditional distribution
$F_{X,Y}(x, y)$	joint distribution
$G(s), G_X(s)$	generating function
I_A	indicator of the event A
J	Jacobian
$M_X(t)$	moment generating function
H, T	head, tail
$N(\mu, \sigma^2)$	normal distribution
$N(t)$	Poisson process
$X, Y, Z, X(\omega)$	random variables
X, Y, W	random vectors
$W(t)$	Wiener process
\mathcal{F}	σ -field or event space
η	probability of extinction
$\chi^2(\cdot)$	chi-squared distribution
$\phi(x)$	standard normal density
$\Phi(x)$	standard normal distribution
x	vector

\mathbf{x}^T	vector transpose
A	matrix
A^T	matrix transpose
$\mu^{(k)}$	fractional moment
μ	mean
μ_i	mean recurrence time
$\boldsymbol{\pi}, \pi_i$	stationary distribution
σ^2	variance
$\rho(X, Y)$	correlation
$\Gamma(t)$	gamma function
Ω	sample space
$\mathbf{E}X$	expectation of X
$\mathbf{E}(X Y)$	conditional expectation
$\mathbf{P}(A)$	probability of A
$\mathbf{P}(A B)$	conditional probability
ϕ	empty set, impossible event
$O(\cdot), o(\cdot)$	order notation

ENVOY

Probability begins and ends with probability.

John Maynard Keynes.

Index

Abbreviations used in this index: *distn* distribution; *eqn* equation; *fn* function; *mgf* moment generating function; *pgf* probability generating function; *pr* process; *rv* random variable; *rw* random walk; *thm* theorem

- absolutely convergent, 22
- absolute probabilities, 403
- absorbing state, 405
- abuse of notation, 309
- accidents and insurance, 72–73
- Acme gadgets, 144–145
- addition rules, 32–34
- ageing and survival, 312–314, 330
- Alea, 14
- α -field, 31
- American call option, 473
- American roulette, 5–6
- AM/GM (arithmetic-geometric means) inequality, 133–134
- ant on square lattice, 94, 112
- aperiodic state, 404
- arcsine law, 216, 262–263, 471–472
- arcsine law for zeros, 471–472
- arithmetic sum, 21
- ark, 87
- astronomy, need for models in, 3
- asymptotic equipartition, 423–424
- attractive events, 56–57
- atypical sequences, 424
- augury, 4
- average, definition, 120. *See also* expectation
- averages, law of, 196–199
- axioms of probability, 32

- Bachelier, L., 438
- backward martingales, 195–196
- bad luck, 385
- balance of life, 382–383
- ballot thm, 112, 189–190
- barriers, retaining, 413
- base rate fallacy, 12
- basic addition rule, 32
- Bayes's thm, 54–55, 64
- Bell's inequality, 229
- Bernoulli, Daniel, 208
- Bernoulli diffusion model, 455
- Bernoulli patterns, 275–276, 459–460
- Bernoulli's urn, 229
- Bernoulli trials, 115, 134, 172. *See also* indicator fns
- Bertrand's paradox, 63–64, 324–325, 346
- beta rv, 334
- biased dice, 7–8, 82, 270–271
- bilateral exponential, 294–295, 297
- binomial
 - distn, 116–118, 138–139
 - factorial moments, 167
 - limits, 149–150
 - pgf, 238, 240, 242–243, 264
 - rv, 134
 - sums, 172, 176–177
 - symmetric, 455
- binomial coefficients, 85, 89
- binomial thms, 21, 91–92, 95–96, 139
- binormal distn, 387–388
- birthdays, 50, 109
- birth-death pr, 465–470
- birth pr
 - birth-death pr, 465–470
 - boys and girls, 8, 73–74, 87
 - disasters, 463–465
 - family planning, 44–45
 - forward eqns, 428–431
 - general, 465–466
- bivariate generating fns, 90
- bivariate normal density, 342, 348–349, 352, 359–360, 373–376, 440
- bivariate rejection, 392
- Black-Scholes formula, 472–473
- bookmaker example, 448–449
- Boole's inequalities, 36, 39, 50
- bounded convergence thm, 201
- boys and girls, 8, 73–74, 87
- branching pr
 - family tree, 243–244
 - geometric, 272–273
 - Markov chains, 397
 - martingales for, 278
 - pgf, 243–244, 264, 271–272, 278
 - total population in, 280
- broads, 82
- Brown, Robert, 438
- Brownian bridge, 441, 444–445, 471
- Brownian motion, 438, 442, 447. *See also* Wiener pr
- Buffon's needle, 377–379

- calculus, 265–267
- call option, 448, 472–473
- Camelot, 111
- cans without labels, 153
- capture-recapture, 216–217
- cardinality (size), 19, 29
- cards, 51–53, 114, 160–161, 178, 179
- Carroll, Lewis, 81–82, 394
- Cauchy-convergent, 200
- Cauchy density, 295–296, 304, 320, 391
- Cauchy distn, 295–296
- Cauchy-Schwarz inequality, 170–171, 211, 352
- central joint moments, 167
- Central Limit Theorem (CLT), 370, 388–389, 394
- central moments, 123, 124–125, 135
- certain events, 26
- chain. *See* Markov chain in
 - continuous time; Markov chain
 - in discrete time
- chance, examples of, 1–3
- change of variables technique, 342–344, 361, 381
- Chapman-Kolmogorov eqns, 402, 426, 428, 439, 449–450
- characteristic fns, 390–391
- Chebyshev-Hermite polynomials, 324
- Chebyshev's inequality, 131–132, 156, 306, 370
- chi-squared density, 297, 299

- closed state, 405
 coats, 223–224
 coincidences, 12–13, 258
 coin tosses
 Bernoulli patterns, 275–276, 459–460
 conditional distn, 129–130
 fair, 190–191
 generating fns, 235–236, 242–243, 278
 geometric distn, 124–125, 175–176
 independence, 57–58
 joint distn, 161–162
 by machine, 389
 Motto, 476
 Murphy's law in, 46–47
 notation, 27
 runs, 103–104, 111, 275–276
 simulations, 301
 visits of a rw, 219
 colouring, 106
 combinations, 86–87
 complacency example, 67–68
 complements, 17–18, 26, 39
 composition, 302
 compound Poisson pr, 462
 conditional density, 310–312, 319, 355–361, 365–366, 373, 440
 conditional distn, 127–130, 135–136, 310–312, 319, 356, 373
 conditional entropy, 214, 405
 conditional expectation, 128–129, 136, 177–183, 204–205, 311, 357, 373
 conditional Gambler's Ruin, 70–72, 180–181, 230–231
 conditional independence, 57–58, 65, 177–183, 205
 conditional mass fns, 116–121, 127–128, 178, 211–212
 conditional mgf, 359–360
 conditional probability, 51–82
 Bayes's thm, 54–55
 conditional distn, 127–130
 definition, 52
 generating fns, 235
 independence and, 57–60
 overview, 51–57
 rv, 127–130
 rw, 189
 recurrence and difference eqns, 60–62
 repellent and attractive events, 56–57
 conditional property of Poisson pr, 365–366
 conditional rw, 189
 conditional variance, 205, 440
 conditional Wiener pr, 440
 conditioning rule, 64, 135–136, 319
 congregations, 210–211
 constant of proportionality, 5
 continuity, in calculus, 265–266
 continuity thm, 251, 308
 continuous partition rule, 361, 373, 376–377
 continuous rv, 287–336. *See also*
 jointly continuous rv
 ageing and survival, 312–314
 conditional distn, 310–312
 density, 291–297
 discrete approximation, 300–301
 expectation, 302–306, 319
 fns of, 297–301, 318–319
 inverse fns, 299–300
 mgf, 306–310, 319
 normal distn, 296, 299, 303, 306–310, 320, 323–324
 random points, 315–318
 step fns, 300
 stochastic ordering, 314–315
 uniform distn, 287–290, 297–299
 continuous set fns, 37
 continuous time. *See* Markov chain
 in continuous time
 convergence, 22, 199–203, 265
 convexity, 132–133
 convolutions, 91, 175, 257, 411
 cooperation, 214–215
 correlation coefficient, 170, 172, 204, 206–207, 352, 373
 countable additivity, 34–35
 countable sets, 27
 countable union, 27
 counting, 83–113
 coin tosses, 103–104
 colouring, 106
 combinations, 86–87, 95
 derangements, 88, 96
 dice example, 84
 first principles, 83–84
 generating fns, 90–93, 95
 Genoese Lottery, 98–99
 identity example, 102–103
 inclusion-exclusion, 87–88
 lottery examples, 98–101
 matching, 107–108, 153, 173
 Ménages problem, 101–102
 permutations, 84–86, 95
 railway trains, 97–98
 recurrence relations, 88–89
 ringing birds, 99–100
 techniques, 93–95
 couples survival, 208–209
 coupling technique, 419–420
 coupon collectors, 92–93, 110, 126–127, 156–157, 166, 386
 covariance, 167–169, 204, 248, 352–353
 craps, 45–46, 59–60, 82
 Crofton, M. W., 315, 317, 376
 crossing the road, 277
 cubes, rw on, 417, 451–452
 cumulant generating fns, 247, 271
 cumulative distn fn, 117–118
 cups and saucers, 42
 current life or age, 382–383
 cutting for the deal, 160–161, 168–169, 178, 179
 dart throws, 26, 115
 decay, 462–463
 decomposition thm, 425
 defective rv, 239, 244
 delayed renewal, 256–258
 de Moivre, Abraham, 15, 20, 137, 334
 de Moivre-Laplace thm, 309, 333–334
 de Moivre's thm, 122
 de Moivre trials, 247–248
 de Morgan's laws, 37–38
 density. *See also* distribution
 beta, 334
 bivariate normal, 342, 348–349, 352, 359–360, 373–376
 calculus, 266
 Cauchy, 295–296, 304, 320, 391
 chi-squared, 297, 299
 conditional, 310–312, 319, 355–361, 365–366, 373, 440
 definition, 291, 318
 expectation and, 304–305, 355–361
 exponential, 292–294, 297–298, 303, 311–312, 314, 320
 gamma, 297, 299, 307, 320, 380–381
 independence and, 344–345, 347
 joint. *See* joint density
 Laplace, 320
 marginal, 341, 356, 371
 mixtures, 297, 318, 335
 mgf and, 306–310
 multinormal, 373–374
 multivariate normal, 373–374, 394
 normal. *See* normal density
 order statistics, 363–364
 Pareto, 303
 standard normal, 296, 299
 sums, products, and quotients, 348–351
 triangular, 228
 trinormal, 373–374, 387–388
 two-sided exponential, 294–295, 297
 uniform. *See* uniform density
 Weibull, 313, 322
 dependence, 169–171, 177–183
 derangements, 88, 96, 113
 derivatives, 266
 detailed balance eqns, 454
 determinism, 11
 "De Vetula," 6, 14

- dice
 cooperation, 214–215
 counting principles and, 84
 craps, 45–46, 59–60, 82
 dodecahedral, 270
 models for, 6, 25–26, 34, 40–41
 regeneration, 255
 sixes, 43–44, 182
 tetrahedral, 270
 unorthodox numbers on,
 269–271
 weighted, 7–8, 82, 270–271
 difference, 27, 39
 differentiable fns, 266, 267
 diffusion models, 454–455
 disasters, 463–465
 discrete approximation, 300–301
 discrete rv, 114
 discrete renewal pr, 256
 disjoint events, 17, 29, 63
 distribution
 beta, 334
 binomial. *See* binomial distn
 binormal, 387–388
 Cauchy, 295–296
 chi-squared, 297, 299
 conditional, 127–130, 135–136,
 310–312, 319, 373
 continuous rv, 287–290
 convergence, 199–200
 in counting, 83–84
 cumulative, 117–118
 current life, 382–383
 defective, 239, 244
 discrete, 114
 excess life, 382–383
 expectation, 302–306
 exponential, 292–294, 297–298,
 303, 311–312, 314, 320
 fns, 117–118
 gamma, 297
 Gaussian. *See* normal distn
 geometric, 124–125, 134, 137,
 217, 292
 hypergeometric, 134
 joint, 158–162, 337–342
 logarithmic, 282
 marginal, 341, 356, 371
 moments, 306
 negative binomial, 117–118, 134
 negative hypergeometric, 217
 normal. *See* normal distn
 Poisson, 117, 130–131, 139–141,
 145, 364–368, 389
 rv, 115–120
 sample space and, 287–288
 sequences of, 130–131
 standard trivariate normal,
 373–374
 stationary, 412–418, 434–435,
 436, 449–450, 454–455
 triangular, 228
 trinormal, 374, 387–388
 two-sided exponential, 294–295,
 297
 uniform, 123, 178, 288, 291,
 338–339, 345–346
 dog bites, 139–141, 143–144
 dogfight example, 68–69
 dominated convergence, 201
 doubly stochastic matrices, 400, 417
 doubly stochastic Poisson pr, 462
 drifting Wiener pr, 442, 468–469
 duelling, 147–149

 Eddington's controversy, 75–76
 eggs, 180, 249–250
 Ehrenfest model for diffusion, 455
 eigenvectors, 413
 Einstein, Albert, 438
 entropy, 150–151, 212–214, 405,
 421
 epidemics, 301–302
 equilibrium, of Markov chains, 413,
 431–436
 estimation of mass fns, 199
 European call option, 448, 472–473
 events
 attractive, 56–57
 certain, 26
 definition, 25, 39
 disjoint, 17, 29, 63
 event space, 31
 impossible, 26
 independence of, 57–60
 recurrent, 256, 258, 457
 repellent and attractive, 56–57
 sequences of, 36–37
 useful identities, 28
 event space, 31
 excess life, 382–383
 expectation, 120–127
 conditional, 128–129, 136,
 177–183, 204–205, 311, 373
 conditional density and, 355–361
 continuous rv, 302–306, 319
 defective, 239, 244
 definition, 120, 135
 densities and, 302–306, 355–361
 fns of rv, 120–123, 205, 318–319
 joint distn, 165–172
 jointly continuous rv, 351–360,
 372
 tail integral, 305
 tail sums, 156
 experiments, 24, 26
 exponential characteristic fns,
 390–391
 exponential density, 292–294,
 297–298, 303, 311–312, 314,
 320
 exponential distn, two-sided,
 294–295, 297
 exponential fns, 91
 exponential generating fns, 90,
 92–93, 95, 100, 308
 exponential limit thm, 23
 exponential variable, von
 Neumann's, 383–385
 exponential Wiener pr, 442–443
 extended addition rules, 32–34, 64
 extended multiplication rule, 64
 extended partition rule, 54

 factorial moments, 123, 125,
 237–238, 246, 253
 failure rates, 312–314, 330
 fair games, 190–196, 448–449
 fair odds, 13
 fair value, 8–9
 falling off a log, 427–428
 false positives, 55
 family, of particles, 243
 family planning, 44–45
 family size, 465–466, 469–470
 α -field, 31
 fingerprints, 152–153
 finite additivity, 34–35
 finite series, 20–21
 first moment, 120. *See also*
 expectation
 first passage times
 of Markov chains, 406–412,
 443–445, 449–450, 459–460
 mean, 406
 preservation at, 410
 of rw, 185, 193, 259
 Wiener pr, 443–444
 fish, 105
 flats, 82
 flies, 61–62
 flowers, 58–59
 F-measurable, 288
 forward eqns, 429, 431–436
 frogs, 80
 functions
 beta, 334
 characteristic, 390–391
 continuous rv, 297–301, 318–319
 cumulant, 247, 271
 differentiable, 266, 267
 exponential, 91
 exponential characteristic,
 390–391
 gamma, 297, 332
 generating. *See* generating fns
 hazard, 312–314, 330
 indicator, 20, 115, 120, 124,
 172–174, 359
 integrable, 266
 inverse, 20, 299–300
 joint mass, 158–159, 162–165,
 171, 206–207, 211–212
 limits of, 265
 logarithmic, 264
 log-survivor, 312–314, 330
 of more than one variable,
 267–268
 rv, 135, 203

- functions (*cont.*)
 reliability, 312–314, 330
 step, 300
 tail generating, 234, 263
 fundamental thm of calculus,
 266–267
- Galileo, 437
 Galton's paradox, 63
 Gambler's Ruin
 conditional, 70–72, 180–181,
 230–231
 generating fns, 268–269
 Markov chains, 419
 rv, 142–143
 as rw, 184–185, 269
 three-handed, 224–226
 gambling, martingales in, 145–146,
 190–196, 222–223, 224–226
 gamma density, 297, 299, 307, 320,
 380–381
 gamma fn, 297, 332
 Gaussian distn. *See* normal distn
 generating fns, 232–286
 applications of, 97–100
 binomial, 238, 240, 242–243, 264
 bivariate, 90
 branching pr, 243–244, 264,
 271–273, 278
 conditional mass fns, 235
 convolutions, 257
 cumulant, 247, 271
 definition, 90, 232
 exponential, 90, 92–93, 95, 100,
 308
 factorial moments, 237–238, 246,
 253
 geometric, 239, 264
 independent rv, 239–244, 249,
 264
 joint, 247–250, 264, 353
 logarithmic, 264
 moment. *See* moment generating
 fns
 moments and, 236–239,
 251–252, 263
 Poisson, 238, 240, 249–250, 264
 rw, 259–263
 regeneration, 254–259
 sequences, 251–254
 sums and random sums, 263
 tail, 234, 263
 trinomial, 283
 uniqueness thm, 90, 237, 263
 genetics, 58–59
 Genoese Lottery, 98–99
 Gentile, Benedetto, 99
 geometric branching, 272–273
 geometric distn
 continuous rv, 292
 discrete rv, 124–125, 134, 137
 lack-of-memory property,
 311–313, 357, 427
 pgf, 239, 264
 sum of, 175–176
 geometric pgf, 239, 264
 geometric sums, 21
 geometric Wiener pr, 442–443
 Gibbs inequality, 151
 goats and cars, 12, 75, 82
 god, as model for chance, 4
 golf, 206–207, 248–249
 goodness-of-fit tests, 141
 guesswork, 141–142
- Hall, Monty, 12, 75, 82
 hazard fn, 312–314, 330
 hazard rate fn, 312–314, 330
 Hermite polynomials, 324
 history of probability, 14–15, 34
 hitting time thm, 187–188, 261–262
 homogeneity, 367
 Hooke, Robert, 437
 hypergeometric rv, 134
- identities, combinatorial, 93–94,
 102–103
 ill-posed questions, 73
 imbedding, 398
 immigrants, 463–465
 impossible events, 26, 39
 inclusion, 27
 inclusion-exclusion inequalities,
 174
 inclusion-exclusion principle, 39,
 87–88, 100
 increments, 203, 220, 366–368
 independence
 conditional, 57–58, 65, 177–183,
 205
 correlation coefficients, 170, 204
 definition, 344
 events, 57–60
 generating fns, 239–244, 249,
 264
 joint distn, 162–165
 jointly continuous rv, 344–347,
 372
 of mean and variance, 354
 pairwise, 58, 65, 250
 independence of sample mean and
 sample variance property, 354
 independence rule, 64
 independent increments property,
 366–368
 indicator fns
 binomial rv, 172
 events, 120, 124
 as inverse fns, 20
 jointly continuous rv, 359
 linear, 172–174
 Markov chains, 421–422
 matching, 173
 simple, 115
 inequalities, 131–134
 AM/GM, 133–134
 basic, 131
 Bell's, 229
 Boole's, 36, 39, 50
 Cauchy-Schwarz, 170–171, 211,
 352
 Chebyshev's, 131–132, 156,
 306, 370
 Gibbs, 151
 inclusion-exclusion, 174
 Jensen's, 133, 231, 306
 Markov, 132
 as sums and products of rv,
 172–177
 infinite series, 22–23
 information, 212–214
 information sources, 397
 inspection paradox, 382–383
 insurance, 72–73
 integrable fns, 266
 integrals, 266
 integrated Wiener pr, 443
 integration, Monte Carlo, 198, 316
 intensity of probability, 313
 intersections, 27
 intuition, 6, 7, 10
 inverse fns, 20, 299–300
 inversion thm, 308
 irreducibility, 404, 434–435
- Jacobian, 343
 Janssen, Hans and Zacharias, 437
 Jensen's inequality, 133, 231, 306
 joint density
 change of variables, 342–344
 definition, 337–338
 distn, 337–342
 expectation, 351–355
 independence, 344–347
 sums, products, and quotients,
 348–351
 triangles, 339–340
 Wiener pr, 439
 joint distn, 158–162, 203,
 337–342
 joint generating fns, 247–250, 264,
 353
 joint lives, 208–209
 jointly continuous rv, 337–395
 change of variables, 342–344,
 361, 381
 conditional density and
 expectation, 355–361
 expectation, 351–355, 372
 independence, 344–347, 372
 joint density and distn, 337–342
 Poisson pr martingales,
 364–368
 sums, products, and quotients,
 348–351
 transformations, 361–364, 372
 joint mass fns, 158–159, 162–165,
 171, 206–207, 211–212
 joint mgf, 353–354, 372–373
 joint moments, 167, 197

- Key Rule
 for conditional distn fn, 356
 for conditional mass fns, 204
 for density, 291, 318
 for discrete rv, 135, 136
 for joint distn, 159, 203
 for jointly continuous rv, 339, 371
 for rv, 116
- Kolmogorov, A., 15
- Kronecker delta, 399
- lack-of-memory property, 311–313, 357, 427
- Laplace, Pierre-Simon de, 11
- Laplace-de Moivre thm, 309, 333–334
- Laplace density, 320
- law of averages, 196–199
- laws
 arcsine, 216, 262–263
 arcsine for zeros, 471–472
 of averages, 196–199
 de Morgan's, 37–38
 Murphy's, 46–47, 474
 strong, of large numbers, 198
 weak, of large numbers, 197–199, 370, 421
- lemons, 53–54
- Lewis Carroll, 81–82, 394
- L'Hôpital's rule, 238, 258
- library books, 417–418
- life
 balance of, 382–383
 current, 382–383
 excess, 382–383
 joint, 208–209
 residual, 314–315
- light bulbs, 277, 382–383
- limits
 binomial, 149–150
 central limit thm, 370, 388–389, 394
 definition of, 21
 exponential, 23
 fns of, 265
 integrals as, 266
 Markov chain, 425
 O - o notation, 309
- Lippeshey, Hans, 437
- lists, in sets, 16–17
- location parameters, 120
- logarithmic distn, 282
- logarithmic generating fns, 264
- log-survivor fn, 312–314, 330
- lottery examples, 27, 36, 98–101, 111, 118–119, 137–138
- lumping states, 399, 474
- machine, 254–255, 389, 432–434
- marginal distn, 341, 356, 371
- marginal mass fns, 159, 162–163, 168, 203, 206, 211–212
- marking animals, 216–217
- Markov chain in continuous time, 436–449. *See also* Wiener pr
 Chapman-Kolmogorov eqns, 402, 449–450
 first passage times, 443–445, 449–450
 forward eqns, 429, 431–436
 irreducibility, 404, 434–435
 Ornstein-Uhlenbeck pr, 442
 stationary distn, 436, 449–450, 454–455
 transition matrix, 454
 Wiener pr definition, 436–437
- Markov chain in discrete time, 396–477
 branching pr, 397
 Chapman-Kolmogorov eqns, 402, 426, 428, 449–450
 with continuous parameter, 425–428
 coupling, 419–420
 decomposition, 425
 definitions, 396–398
 diffusion models, 454–455
 doubly stochastic, 417
 entropy of, 405, 421
 equilibrium, 413, 431–436
 finite, 397
 first passage times, 406–412, 443–445, 449–450, 459–460
 forward eqns, 431–436
 imbedded, 398
 indicator fns, 422
 irreducibility, 404, 434–435
 limit thm for, 425
 in the long run, 418–425
 lumping states, 399
 nonhomogeneous, 475
 notation, 398
 paths, 408
 persistence, 457–458
 Poisson pr, 461–462
 recurrence time, 424–425
 regular, 404, 415
 renewal sequence, 456–457
 reversible chains, 453–454
 simple rw, 397, 401, 411, 419
 stationary distn, 412–418, 434–435, 436, 449–450, 454–455
 transition probabilities, 400–406, 408, 410, 426, 449
 two state chain, 403
 weak law of large numbers, 421
- Markov inequality, 132
- Markov property, 396–400
 branching property, 397
 definition, 396–398
 at first passage times, 410
 jointly continuous rv, 383
 lumping states and, 399, 474
 messages, 397
 Poisson pr and, 427–431
- simple rw, 397
 strong, 410, 444
 Wiener pr, 437
- Markov sources, 398, 405–406, 423–424
- Markov time, 474
- martingales, 190–196
 backward, 195–196
 bounded, 225
 branching, 278
 coin toss, 190–191
 conditional expectation, 191
 convergence, 202
 definition, 190
 Doob, 192
 with finite variance, 220–221
 in gambling, 145–146, 190–196, 222–223, 224–226
 Markov chain, 445–448, 469–470
 matching, 223–224
 optional stopping, 194, 201, 220, 231, 368, 447
 option-pricing, 447–448
 options, 192
 orthogonal increments, 203, 220
 Poisson, 367–368, 389
 roulette, 145–146, 193
 sequences, 196
 simple rw, 221–222
 stopping times, 192–194, 220
 submartingales, 195–196, 202
 supermartingales, 195–196, 202, 224
 Wald's eqn, 195
 Wald's identity, 279–280
 Wiener pr, 445–446, 470
- mass fns
 conditional, 116–121, 127–128, 178, 211–212
 estimation of, 199
 joint, 158–159, 162–164, 171, 193, 206–207, 211–212
 marginal, 159, 162–163, 168, 203, 211–212
 probability, 33–34
- mass metaphor, 33–34
- matching
 coincidences, 107–108
 expectations, 231
 factorial moments, 153
 generating fns, 252–253
 indicators, 173
 martingales, 223–224
 Poisson distn, 130–131
- mathematical models, advantages of, 4–5
- mathematics review, 15–23
 fns, 19–20
 indicator fns, 20
 inverse fns, 20
 limits, 21–22
 notation, 15–16
 series and sums, 20–23

- mathematics review (*cont.*)
 sets, 16–19
 size, 19
 Venn diagrams, 17–18
 matrix
 doubly stochastic, 400, 417
 positive, 400
 stochastic, 400, 412, 417
 substochastic, 400
 transition, 400, 402, 415, 454
 maxima from minima, 385–386
 mazes, 254
 mean, 120, 256, 406. *See also*
 expectation
 mean-square convergence, 199–200
 measurability, 288
 medians, 120, 138
 Ménages problem, 101–102, 112,
 130–131
 messages, 397
 Mill's ratio, 323
 minima for maxima, 385–386
 misprints, 138–139
 mixtures, 297, 318, 335
 models
 advantages of, 4
 American roulette, 5–6
 dice, 6, 25–26, 34, 40–41
 diffusion, 454–455
 purposes of, 3–4
 modulus of elasticity, 335
 molecules, 394
 moments. *See also* moment
 generating fns
 central, 123, 125, 135
 central joint, 167
 covariance, 204
 definitions, 123, 135
 density and, 306
 factorial, 123, 125, 237–238,
 246, 253
 generating fns and, 236–239
 joint, 167, 197, 353–354
 second, 124–125
 moment generating fns
 conditional, 359–360
 continuous rv, 306–310, 319
 cumulants, 247
 definition, 245–247, 263
 densities and, 306–310, 359–360
 factorial moments, 246
 joint, 264, 353–354, 372–373
 sequences, 251–252
 monotone convergence thm,
 201–202
 Monte Carlo integration, 198, 316
 Monty Hall problem, 12, 75, 82
 mortality fn, 312–314, 330
 Motto, 476
 multinomial coefficients, 85,
 93
 multinomial thm, 21, 93, 96
 multinormal distn, 373–374
- multiplication rule, 64
 multivariate normal density,
 373–374, 394
 Murphy's Law, 46–47, 474
- needle, Buffon's, 377–379
 negative binomial distn, 117–118,
 134
 negative binomial thm, 22, 117
 negative hypergeometric distn, 217
 nonhomogeneous Markov chain,
 475
 nonhomogeneous rw, 414–415
 normal density. *See also* normal
 distn
 continuous rv, 296, 299, 303,
 306–308, 320, 323–324
 de Moivre-Laplace thm and, 310
 jointly continuous rv, 346–347
 normal distn
 binormal, 387–388
 bivariate, 342, 348–349, 352,
 359–360, 373–376, 440
 central limit thm, 370, 388–389,
 394
 correlation, 373, 440
 covariance, 373
 linear transform, 362–363
 multinormal, 373–374
 multivariate, 373–374, 394
 standard, 296, 299
 standard trivariate, 374
 trinormal, 373–374, 387–388
 notation
 abuse of, 309
 for events, 27, 38–39
 Markov chains, 398
 mathematical, 15–16
 O-*o*, 309
 for outcomes, 26–27, 38–39
 null state, 425
- occupancy, 274–275
 odds, 13
O-*o* notation, 309
 opinion polls, 27
 optional skipping, 223
 optional starting, 223
 optional stopping, 194, 201, 223,
 231, 446–447
 option pricing, 472–473
 option-pricing martingale, 447–448
 ordered pairs, notation of, 27
 ordering, 219–220
 order statistics, 363–364
 Ornstein-Uhlenbeck pr, 442
 orthogonal increments, 203, 220
 orthogonal variables, 197
 outcomes, 24–25, 27
- pairwise independence, 58, 65,
 250
 Pap smears, 69–70
- paradoxes
 Bertrand's Box, 63–64
 Carroll's, 81–82
 Galton's, 63
 inspection, 382–383
 prisoners', 81
 sampling, 211
 Simpson's, 81
 Pareto density, 303
 partition rule, 53–54, 64, 97–98,
 181, 204
 partitions, 18, 376–377
 Pascal's triangle, 89
 path, 184
 pay-off odds, 8–9, 13
 periodic state, 404
 permutations, 84–86, 88, 95, 97–98
 Perrin, Jean-Baptiste, 438
 persistence, 457–458
 persistent renewal pr, 256–257,
 259–260, 457–458
peçsoi, 14
 pirated copies, 329–330
 Poisson distn
 dogbite example, 139–141
 martingales, 364–368, 389
 pgf, 238, 240, 249–250, 264
 rv, 117, 145
 sequences, 130–131
 Poisson pr, 364–368, 427–431,
 461–462
 poker, 51–53, 114
 Polya's Urn, 66–67, 231
 pooling, 81
 positive matrices, 400
 postmen, 139–141, 143–144
 power series, 233
 prisons, 302
 prisoners' paradox, 81
 probability
 American roulette, 5–6
 axioms of, 32
 conditional, 51–57. *See also*
 conditional probability
 continuity, 251, 308
 convergence, 199–200
 definition, 25
 history of, 14–15, 34
 importance of the long run, 7–8
 intuition on, 6, 7, 10
 mass metaphor, 33–34
 as measure of expected value,
 33
 notation, 38–39
 as a number, 3
 pay-offs, 8–9
 reality and, 11–12
 as relative frequency, 33
 symmetry in, 5–7
 transition. *See* transition
 probabilities
 probability distn, 32, 39
 probability fns, 34–35, 36, 39

- probability generating fns, 232–286, 263. *See also* generating fns
- probability mass fns. *See* mass fns
- probability statements, 25
- Problem of the Points, 14–15
- propagation, 211–213
- proper rv, 116
- pseudorandom numbers, 301–302
- pull-through property, 181, 205
- quantum theory, 12, 229
- queue, 435–436
- quotients, 348–351
- railway trains, 97–98
- random points, 315–318
- random pr, 426
- random sample. *See* normal distn
- random variables, 114–157
- Bernoulli trials, 115, 134, 172
 - beta, 334
 - binomial, 116–118, 134, 138–139, 149–150, 167, 172
 - binormal, 387–388
 - Cauchy, 295–296
 - chi-squared, 297, 299
 - conditional distn, 127–130, 135–136
 - conditional expectation, 128–129, 136
 - continuous, 287–336. *See also*
 - continuous rv
 - defective, 239, 244
 - discrete, 114–118, 134
 - distn, 115–120
 - expectation, 120–127, 135, 204
 - F-measurable, 288
 - geometric distn, 125, 134, 137–138, 175–176
 - hypergeometric, 134
 - indicators, 115, 120, 124, 172–174
 - inequalities, 131–134, 177
 - joint density and distn, 337–342
 - Key Rule, 116, 135, 136
 - medians, 120
 - moments, 123–125, 135
 - multinomial, 85, 93
 - negative binomial, 117–118, 134
 - notation, 114
 - orthogonal, 197
 - Poisson, 117, 130–131, 134, 139–141, 145
 - products of, 172–177
 - proper, 116
 - quotients, 348–351
 - sequences of distns, 130–131
 - simple, 115
 - simulation of, 301–302
 - sums, 124, 172–177
 - symmetric, 197
 - tails, 124, 156
 - trinormal, 373–374, 387–388
 - uncorrelated, 197
 - uniform, 123, 134, 178
 - variance, 123–124
- random vectors, 158–231
- conditional expectation, 177–183
 - convergence, 199–203
 - correlation coefficient, 170
 - expectation, 165–172
 - inclusion-exclusion inequalities, 174
 - independence, 162–165
 - inequalities, 177
 - joint distn, 158–162
 - law of averages, 196–199
 - martingales, 190–196
 - simple rw, 183–190
 - sums and products of rv, 172–177
- random walk
- arcsine law for leads, 262–263
 - ballot thm, 189–190
 - Brownian motion as, 438
 - coincidences, 258
 - conditioned, 189
 - on cubes, 417, 451–452
 - first passage times, 185, 193, 259
 - generating fns, 259–263
 - hitting time thm, 187–188
 - imbedded, 398
 - Markov chains, 397, 401, 411, 419
 - martingales, 221–222
 - nonhomogeneous, 414–415
 - path of, 184
 - persistent, 257, 259–260
 - recurrence time, 424–425
 - reflection principle, 445
 - with retaining barriers, 413
 - simple, 183–190, 221–222
 - simple symmetric, 183, 188–189, 222, 259
 - truncated, 261–262
 - visits of, 218–219
- range, 39
- rare diseases, 55
- reality, 11–12
- recapture, 216–217
- recurrence and difference eqns, 60–62, 89, 90
- recurrence relations, 88–89, 90, 107–108
- recurrence time, 185, 256–257, 406–408, 424–425
- recurrent event pr, 256, 258, 457
- recurrent null state, 425
- reflected Wiener pr, 443
- reflecting barriers, 413
- reflection principle, 187–188, 445
- regeneration, 254–259
- regular chain, 404, 415
- rejection method, 381–382
- reliability fn, 312–314, 330
- rencontres, 107–108
- renewal pr
- chains, 456–457
 - current life, 382–383
 - delayed, 256–258
 - discrete, 256
 - excess life, 382–383
 - machine, 254–255
 - persistent, 256–257, 259, 457–458
 - stationary, 258–259
 - transient, 256–257
- renewal sequence, 456–457
- repellent events, 56–57
- residual life, 314–315
- retaining barriers, 413
- reversible chains, 453–454
- Riemann's hypothesis, 9
- ringing birds, 99–100
- Robbins' improved formula, 96
- roulette, 145–146
- royal flush, 51–52
- Royal Oak Lottery, 137–138
- ruin. *See* Gambler's Ruin
- rules
- addition, 32–34
 - conditional independence, 65, 205
 - conditioning, 64, 135–136, 319
 - continuous partition, 361, 373
 - extended multiplication, 64
 - L'Hôpital's, 238, 258
 - partition, 53–54, 64, 97–98, 181, 204
 - in sets, 16–17
- runs, 103–104, 111, 275–276
- sample size, 368–369
- sample space, 24, 26, 287–288
- sampling, 398
- sampling paradox, 211
- searching, 146–147
- second central moment, 125
- second derivative, 266
- second factorial moment, 125
- second moment, 124–125
- sequences, 196, 251–254, 456–457
- sequences of distns, 130–131
- sequences of events, 36–37
- series, 20–23
- sets, 16–19
- combining, 17
 - continuous set fns, 37
 - countable, 27
 - rules and lists in, 16–17
 - size, 19, 29
 - Venn diagrams, 17–18, 30, 37–38
- sharps, 82
- Shelmikedmu, 29–31
- shower problem, 383
- simple rv, 115
- Simpson's paradox, 81
- simulation of rv, 301–302, 316–317, 392–393

- sixes, 43–44, 182
- size, 19, 29, 368–369
- smears, 69–70
- sortilege, 2, 4
- space
 - event, 31
 - sample, 24, 26, 287–288
 - state, 396–397, 401, 480
- sport, 58
- standard trivariate normal distn, 374
- state
 - absorbing, 405
 - closed set, 405
 - irreducible, 404, 434–435
 - Markov two state chain, 403
 - null, 425
 - periodic, 404
 - recurrent, 425
 - transient, 425
- state space, 396–397, 401, 425
- stationary distn, 412–418, 434–435, 436, 449–450, 454–455
- stationary increments property, 367
- stationary renewal, 258–259
- step fns, 300
- Stirling's formula, 96, 156, 216, 332–334, 388–389
- stochastic matrices, 400, 412, 417
- stochastic ordering, 314–315
- stochastic Poisson pr, 462
- stock control, 326–327
- stopping, optional, 194, 201, 223, 231, 446–447
- stopping times, 192–194, 367–368, 444
- strange but true, 215–216
- strong laws of large numbers, 198
- strong Markov property, 410, 444
- subjective probability, 9–10
- submartingales, 195–196, 202
- substochastic matrices, 400
- sudden death, 65–66
- sums
 - arithmetic, 21
 - binomial, 176–177
 - density, 348–351
 - geometric, 21, 175–176
 - random vectors, 204
 - tail, 124, 156
- supermartingales, 195–196, 202, 224
- survival, 208–209, 312–314, 330, 401, 418–419
- symmetric difference, 27
- symmetric rw, 183, 188–189, 222, 259
- symmetry, 5–7, 197, 377
- tail generating fns, 234, 263
- tail integral thm, 305
- tail sums, 124, 156
- targets, 379–380
- Taylor's Thm, 90
- telecommunications, 397
- telescope invention, 437
- tennis, 209–210
- tessera, 14
- three-handed gambler's ruin, 224–226
- toll booths, 302
- tournament, 209–210
- tower property, 181, 205
- transformations, 342–344, 361–364, 372
- transient renewal pr, 256–257
- transient state, 425
- transition matrix, 400, 402, 415, 454
- transition probabilities
 - in branching pr, 397
 - definition, 396
 - in Markov chains, 400–406, 408, 426, 449, 461
 - n -step, 401–403, 449
 - in the Poisson pr, 429
 - in simple rw, 397
- trials, Bernoulli, 115, 134, 172.
 - See also* indicator fns
- trials, de Moivre, 247–248
- triangles, 330–332, 339–340
- triangular distn, 228
- trinomial generating fn, 247
- trinomial distn, 373–374, 387–388
- truncated rw, 261–262
- truncation, 474
- two-sided exponential distn, 294–295, 297
- two state chain, 403
- typical sequences, 424
- uncertainty, 150–151, 214, 405
- uncertainty principle, 11
- uncorrelated variables, 197
- unfair odds, 13
- uniform density
 - definition of, 291–293
 - fns of rv, 297–299, 303, 321–322
 - jointly continuous rv, 356
 - mgf, 307
- uniform distn, 123, 338–339, 345–346
- union, of sets, 17, 27
- uniqueness thm, 90, 237, 263
- unit cube, 390
- unordered selection, 86–87
- urns
 - Bernoulli's, 229
 - models, 41–42, 66–67, 95, 118–119
 - Polya's, 66–67, 231
- van Leeuwenhoek, Antoine, 437
- variance
 - conditional, 205, 440
 - independence of, 354
 - for rv, 123–124, 135
 - sums, 180
 - uniform distn, 123–124
- Venn diagrams, 17–18, 30, 37–38
- "De Vetula," 6, 14
- visas, 327–329
- visits, 218–219
- von Neumann's exponential variable, 383–385
- waiting times, 312–314
- Wald's eqn, 195
- Wald's identity, 279–280
- Waring's thm, 274–275
- weak law of large numbers (WLLN), 197–199, 370, 421
- weather examples, 4, 13–14, 82, 411–412
- Weibull density, 313, 322
- Wiener, N., 438
- Wiener pr, 436–449
 - arcsine law for zeros, 471–472
 - barriers, 413
 - Brownian motion, 438, 442, 447
 - Chapman-Kolmogorov eqns, 439, 449–450
 - conditioning, 440
 - definition, 436–437
 - drifting, 442, 468–469
 - exiting a strip, 470
 - exponential, 442–443
 - first passage time, 443–445
 - integrated, 443
 - joint densities, 439
 - martingales, 445–446, 470
 - Ornstein-Uhlenbeck pr, 442
 - reflected, 443, 445
 - sample-path properties, 438–439
 - scaling, 442
 - standard, 437
 - stopping times, 444, 447
 - tied-down, 441
- wyvern, 87
- zeros, arcsine law for, 471–472