

THE THEORY OF PROBABILITY

EXPLORATIONS AND APPLICATIONS



Santosh S. Venkatesh

CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9781107024472

THE THEORY OF PROBABILITY

From classical foundations to advanced modern theory, this self-contained and comprehensive guide to probability weaves together mathematical proofs, historical context, and richly detailed illustrative applications.

A theorem discovery approach is used throughout, setting each proof within its historical setting, and is accompanied by a consistent emphasis on elementary methods of proof. Each topic is presented in a modular framework, combining fundamental concepts with worked examples, problems and digressions which, although mathematically rigorous, require no specialised or advanced mathematical background.

Augmenting this core material are over 80 richly embellished practical applications of probability theory, drawn from a broad spectrum of areas both classical and modern, each tailor-made to illustrate the magnificent scope of the formal results. Over 500 homework problems and more than 250 worked examples are included.

Providing a solid grounding in practical probability, without sacrificing mathematical rigour or historical richness, this insightful book is a fascinating reference, and essential resource, for all engineers, computer scientists and mathematicians.

SANTOSH S. VENKATESH is an Associate Professor of Electrical and Systems Engineering at the University of Pennsylvania, whose research interests include probability, information, communication and learning theory, pattern recognition, computational neuroscience, epidemiology and computer security. He is a member of the David Mahoney Institute for Neurological Sciences, and has been awarded the Lindback Award for Distinguished Teaching.

‘This is a gentle and rich book that is a delight to read. Gentleness comes from the attention to detail; few readers will ever find themselves “stuck” on any steps of the derivations or proofs. Richness comes from the many examples and historical anecdotes that support the central themes. The text will support courses of many styles and it is especially attractive for self-guided study.’

J. J. Michael Steele, University of Pennsylvania

‘This book does an excellent job of covering the basic material for a first course in the theory of probability. It is notable for the entertaining coverage of many interesting examples, several of which give a taste of significant fields where the subject is applied.’

Venkat Anantharam, University of California, Berkeley

‘This book presents one of the most refreshing treatments of the theory of probability. By providing excellent coverage with both intuition and rigor, together with engaging examples and applications, the book presents a wonderfully readable and thorough introduction to this important subject.’

Sanjeev Kulkarni, Princeton University

THE THEORY OF PROBABILITY

SANTOSH S. VENKATESH

University of Pennsylvania



CAMBRIDGE
UNIVERSITY PRESS

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

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

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

www.cambridge.org
Information on this title: www.cambridge.org/9781107024472

© Cambridge University Press 2013

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

First published 2013

Printed and bound in the United Kingdom by the MPG Books Group

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

ISBN 978-1-107-02447-2 Hardback

Additional resources for this publication at www.cambridge.org/venkatesh

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

for

ces, bes, bessie,

mummyyummyyum, mozer,

muti, **cecily stewart venkatesh**, anil,

boneil, banana peel, studebaker, suffering

succotash, anna, banana, baby dinosaur, bumpus,

pompous, boom boom music, twelve ways of biscuit,

booboo, binjammin, rumpus, gumpus, grumpus,

goofy, goofus, **anil benjamin venkatesh**, chachu, achu,

chach, ach, uch, uchie, sweetie, honey, bachu, joy, achoo

tree, muggle artefacts, anna did it, bump, lump, sump,

chachubachubach, stormcloud, jammie, sara, baby

tyrannosaurus, simmy, akka, **saraswathi joy venkatesh**,

vid, id, did, the kid, vidya, viddy, viddydid, viddydid-

dydid, little viddles, viddlesticks, fiddlesticks, wun-

kles, winkie, wunks, winks, pooh, pooh sticks,

imposterous, impostorous, ginormous, broom

closet, hobbit, gump, rump, peewee,

googoogaga, **vidya margaret**

venkatesh

who put up with a frequently distracted and unlovable husband and father.

Contents

Preface	page	xv
A ELEMENTS		
I Probability Spaces		3
1 From early beginnings to a model theory		3
2 Chance experiments		5
3 The sample space		9
4 Sets and operations on sets		12
5 The algebra of events		14
6 The probability measure		16
7 Probabilities in simple cases		19
8 Generated σ -algebras, Borel sets		24
9 A little point set topology		26
10 Problems		30
II Conditional Probability		35
1 Chance domains with side information		35
2 <i>Gender bias? Simpson's paradox</i>		40
3 The theorem of total probability		42
4 <i>Le problème des rencontres, matchings</i>		47
5 <i>Pólya's urn scheme, spread of contagion</i>		49
6 <i>The Ehrenfest model of diffusion</i>		52
7 Bayes's rule for events, the MAP principle		56
8 <i>Laplace's law of succession</i>		59
9 <i>Back to the future, the Copernican principle</i>		61
10 Ambiguous communication		64
11 Problems		66
III A First Look at Independence		71
1 A rule of products		71
2 <i>What price intuition?</i>		74

Contents

3	<i>An application in genetics, Hardy's law</i>	77
4	Independent trials	81
5	<i>Independent families, Dynkin's π-λ theorem</i>	87
6	<i>Problems</i>	90
IV	Probability Sieves	93
1	Inclusion and exclusion	93
2	<i>The sieve of Eratosthenes</i>	99
3	<i>On trees and a formula of Cayley</i>	102
4	Boole's inequality, the Borel–Cantelli lemmas	106
5	<i>Applications in Ramsey theory</i>	109
6	Bonferroni's inequalities, Poisson approximation	113
7	<i>Applications in random graphs, isolation</i>	119
8	<i>Connectivity, from feudal states to empire</i>	121
9	Sieves, the Lovász local lemma	125
10	<i>Return to Ramsey theory</i>	130
11	<i>Latin transversals and a conjecture of Euler</i>	131
12	<i>Problems</i>	135
V	Numbers Play a Game of Chance	139
1	A formula of Viète	139
2	Binary digits, Rademacher functions	141
3	The independence of the binary digits	144
4	The link to coin tossing	148
5	The binomial makes an appearance	150
6	An inequality of Chebyshev	153
7	Borel discovers numbers are normal	155
8	<i>Problems</i>	159
VI	The Normal Law	163
1	One curve to rule them all	163
2	A little Fourier theory I	166
3	A little Fourier theory II	172
4	An idea of Markov	177
5	Lévy suggests a thin sandwich, de Moivre redux	181
6	<i>A local limit theorem</i>	184
7	<i>Large deviations</i>	189
8	<i>The limits of wireless cohabitation</i>	191
9	<i>When memory fails</i>	193
10	<i>Problems</i>	195
VII	Probabilities on the Real Line	197
1	Arithmetic distributions	197
2	Lattice distributions	200

3	Towards the continuum	204
4	Densities in one dimension	206
5	Densities in two and more dimensions	210
6	Randomisation, regression	217
7	<i>How well can we estimate?</i>	221
8	<i>Galton on the heredity of height</i>	223
9	Rotation, shear, and polar transformations	227
10	Sums and products	230
11	<i>Problems</i>	232
VIII	The Bernoulli Schema	235
1	Bernoulli trials	235
2	The binomial distribution	237
3	<i>On the efficacy of polls</i>	239
4	<i>The simple random walk</i>	244
5	<i>The arc sine laws, will a random walk return?</i>	247
6	Law of small numbers, the Poisson distribution	252
7	Waiting time distributions	256
8	<i>Run lengths, quality of dyadic approximation</i>	262
9	<i>The curious case of the tennis rankings</i>	264
10	Population size, the hypergeometric distribution	268
11	<i>Problems</i>	271
IX	The Essence of Randomness	275
1	The uniform density, a convolution formula	276
2	<i>Spacings, a covering problem</i>	280
3	<i>Lord Rayleigh's random flights</i>	284
4	<i>M. Poincaré joue à la roulette</i>	288
5	Memoryless variables, the exponential density	292
6	<i>Poisson ensembles</i>	294
7	<i>Waiting times, the Poisson process</i>	297
8	Densities arising in queuing theory	303
9	Densities arising in fluctuation theory	305
10	Heavy-tailed densities, self-similarity	307
11	<i>Problems</i>	310
X	The Coda of the Normal	315
1	The normal density	315
2	Squared normals, the chi-squared density	319
3	A little linear algebra	320
4	The multivariate normal	324
5	<i>An application in statistical estimation</i>	330
6	<i>Echoes from Venus</i>	337
7	<i>The strange case of independence via mixing</i>	341

Contents

8	<i>A continuous, nowhere differentiable function</i>	346
9	<i>Brownian motion, from phenomena to models</i>	348
10	<i>The Haar system, a curious identity</i>	352
11	<i>A bare hands construction</i>	355
12	<i>The paths of Brownian motion are very kinky</i>	360
13	<i>Problems</i>	363
B FOUNDATIONS		
XI	Distribution Functions and Measure	369
1	Distribution functions	370
2	Measure and its completion	372
3	Lebesgue measure, countable sets	375
4	<i>A measure on a ring</i>	380
5	<i>From measure to outer measure, and back</i>	384
6	<i>Problems</i>	391
XII	Random Variables	393
1	Measurable maps	393
2	The induced measure	397
3	Discrete distributions	399
4	Continuous distributions	402
5	Modes of convergence	406
6	Baire functions, coordinate transformations	409
7	Two and more dimensions	411
8	Independence, product measures	415
9	<i>Do independent variables exist?</i>	420
10	<i>Remote events are either certain or impossible</i>	422
11	<i>Problems</i>	424
XIII	Great Expectations	427
1	Measures of central tendency	427
2	Simple expectations	429
3	Expectations unveiled	433
4	Approximation, monotone convergence	438
5	Arabesques of additivity	446
6	<i>Applications of additivity</i>	451
7	<i>The expected complexity of Quicksort</i>	455
8	Expectation in the limit, dominated convergence	459
9	<i>Problems</i>	462
XIV	Variations on a Theme of Integration	465
1	UTILE ERIT SCRIBIT \int PRO OMNIA	465

2	Change of variable, moments, correlation	471
3	Inequalities via convexity	478
4	L^p -spaces	482
5	Iterated integrals, a cautionary example	485
6	<i>The volume of an n-dimensional ball</i>	492
7	<i>The asymptotics of the gamma function</i>	494
8	<i>A question from antiquity</i>	496
9	<i>How fast can we communicate?</i>	500
10	Convolution, symmetrisation	506
11	<i>Labeyrie ponders the diameter of stars</i>	511
12	<i>Problems</i>	516
XV	Laplace Transforms	523
1	The transform of a distribution	523
2	Extensions	529
3	The renewal equation and process	532
4	<i>Gaps in the Poisson process</i>	536
5	<i>Collective risk and the probability of ruin</i>	538
6	<i>The queuing process</i>	542
7	<i>Ladder indices and a combinatorial digression</i>	546
8	<i>The amazing properties of fluctuations</i>	550
9	<i>Pólya walks the walk</i>	555
10	<i>Problems</i>	557
XVI	The Law of Large Numbers	561
1	Chebyshev's inequality, reprise	561
2	Khinchin's law of large numbers	563
3	<i>A physicist draws inspiration from Monte Carlo</i>	566
4	<i>Triangles and cliques in random graphs</i>	568
5	<i>A gem of Weierstrass</i>	571
6	<i>Some number-theoretic sums</i>	574
7	<i>The dance of the primes</i>	582
8	<i>Fair games, the St. Petersburg paradox</i>	585
9	Kolmogorov's law of large numbers	587
10	<i>Convergence of series with random signs</i>	593
11	<i>Uniform convergence per Glivenko and Cantelli</i>	595
12	<i>What can be learnt per Vapnik and Chervonenkis</i>	599
13	<i>Problems</i>	604
XVII	From Inequalities to Concentration	609
1	Exponential inequalities	609
2	<i>Unreliable transcription, reliable replication</i>	614
3	Concentration, the Gromov–Milman formulation	616
4	Talagrand views a distance	619

Contents

5	The power of induction	625
6	<i>Sharpening, or the importance of convexity</i>	630
7	<i>The bin-packing problem</i>	633
8	<i>The longest increasing subsequence</i>	636
9	<i>Hilbert fills space with a curve</i>	638
10	<i>The problem of the travelling salesman</i>	641
11	<i>Problems</i>	646
XVIII Poisson Approximation		651
1	A characterisation of the Poisson	652
2	The Stein–Chen method	656
3	Bounds from Stein’s equation	657
4	Sums of indicators	660
5	The local method, dependency graphs	663
6	<i>Triangles and cliques in random graphs, reprise</i>	665
7	Pervasive dependence, the method of coupling	668
8	<i>Matchings, ménages, permutations</i>	672
9	<i>Spacings and mosaics</i>	679
10	<i>Problems</i>	685
XIX Convergence in Law, Selection Theorems		689
1	Vague convergence	689
2	An equivalence theorem	692
3	Convolutional operators	695
4	<i>An inversion theorem for characteristic functions</i>	698
5	Vector spaces, semigroups	700
6	A selection theorem	705
7	<i>Two by Bernstein</i>	709
8	<i>Equidistributed numbers, from Kronecker to Weyl</i>	712
9	<i>Walking around the circle</i>	715
10	<i>Problems</i>	717
XX Normal Approximation		719
1	Identical distributions, the basic limit theorem	719
2	<i>The value of a third moment</i>	724
3	<i>Stein’s method</i>	730
4	<i>Berry–Esseen revisited</i>	733
5	Varying distributions, triangular arrays	737
6	<i>The coupon collector</i>	742
7	<i>On the number of cycles</i>	745
8	Many dimensions	747
9	<i>Random walks, random flights</i>	751
10	<i>A test statistic for aberrant counts</i>	753
11	<i>A chi-squared test</i>	759

12	<i>The strange case of Sir Cyril Burt, psychologist</i>	763
13	<i>Problems</i>	767

C APPENDIX

XXI	Sequences, Functions, Spaces	771
1	Sequences of real numbers	771
2	Continuous functions	776
3	Some L^2 function theory	783
	Index	789

Preface

GENTLE READER: Henry Fielding begins his great comic novel *Tom Jones* with these words.

An author ought to consider himself, not as a gentleman who gives a private or eleemosynary treat, but rather as one who keeps a public ordinary, at which all persons are welcome for their money. [...] Men who pay for what they eat, will insist on gratifying their palates, however nice and even whimsical these may prove; and if every thing is not agreeable to their taste, will challenge a right to censure, to abuse, and to d—n their dinner without controul.

To prevent therefore giving offence to their customers by any such disappointment, it hath been usual, with the honest and well-meaning host, to provide a bill of fare, which all persons may peruse at their first entrance into the house; and, having thence acquainted themselves with the entertainment which they may expect, may either stay and regale with what is provided for them, or may depart to some other ordinary better accommodated to their taste.

To take a hint from these honest victuallers, as Fielding did, it strikes me therefore that I should at once and without delay explain my motivations for writing this book and what the reader may reasonably hope to find in it. To the expert reader who finds a discursive prolegomenon irritating, I apologise. There have been so many worthy and beautiful books published on the subject of probability that any new entry must needs perhaps make a case for what is being added to the canon.

THE PAST IS PROLOGUE: The subject of chance is rich in tradition and history. The study of games of chance paved the way for a theory of probability, the nascent science of which begot divers applications, which in turn led to more theory, and yet more applications. This fecund interplay of theory and application is one of the distinguishing features of the subject. It is too much to hope to cover all of the facets of this interaction within the covers of one volume—or indeed many such volumes—and I shall look to the history for guidance.

A central thread running through the theory right from its inceptions in antiquity is the concept peculiar to chance of “statistical independence”. This is the notion that rescues probability from being merely a fragrant by-water of the general theory of measure. To be sure one can articulate the abstract idea of “independent functions” but it appears to have little traction in measure outside of the realm of probability where it not only has a profound impact on the theory but has a peculiarly powerful appeal to intuition.

Historically, the concept of statistical independence was identified first with independent trials in games of chance. Formulations of this principle led to most of the classical results in the theory of probability from the seventeenth century onwards to the early portion of the twentieth century. But the theme is far from exhausted: the last quarter of the twentieth century has seen the serendipitous emergence of new, hugely profitable directions of inquiry on deep and unsuspected aspects of independence at the very heart of probability.

The chronological summary of these new directions that I have included below naturally cannot in its brevity do full justice to all the actors who have helped expand the field. Without pretending to completeness it is intended for the expert reader who may appreciate a quick overview of the general tendency of these results and their connections to the earlier history.

- ◆ Vapnik and Chervonenkis’s beautiful investigation of uniform convergence in 1968 expanded hugely on Glivenko and Cantelli’s classical results dating to 1933. This work spurred the development of empirical process theory and served as an impetus for the burgeoning science of machine learning.
- ◆ Stein unveiled his method of approximate computation of expectations in 1970. The method sketched a subtle and fundamentally different approach to the ubiquitous central limit theorem which dates back to de Moivre in 1733. While it was only slowly that the novelty and genuine power of the idea came to be appreciated, Stein’s method has not only placed the classical canon in an inviting new light, but has opened new doors in the investigation of central tendency.
- ◆ The application of Stein’s method to Poisson approximation was fleshed out by Chen in 1976 and breathed new life into the theory sparked by Poisson’s approximation to the binomial in 1837. The theory that has emerged has provided flexible and powerful new tools in the analysis of rare events, extrema, and exceedances.
- ◆ The Lovász local lemma appeared in 1975 and provided a subtle view of the classical probability sieves used by de Montmort in 1708 and whose provenance goes as far back as the number-theoretic sieves known to the Greeks. It is hard to overstate the abiding impact the local lemma has had;

it and the related sieve arguments that it engendered are now a staple of combinatorial models.

- ◆ The idea that the phenomenon of concentration of measure is very pervasive began to gain traction in the 1980s through the efforts of Gromov and Milman. Talagrand’s stunning paper of 1995 placed an exclamation point on the idea. It is most satisfying that this powerful idea constitutes a vast extension of scope of perhaps the most intuitive and oldest idea in probability—the law of large numbers.

These newer developments all in one way or the other expand on the central idea of independence and, to the great pleasure of this author, connect back, as I have indicated, to some of the oldest themes in probability. In close to twenty five years of teaching and stuttering attempts at writing at the University of Pennsylvania and visiting stints at the California Institute of Technology and the Helsinki University of Technology I have attempted to connect these themes from different perspectives, the story being modulated by the developments that were occurring as I was teaching and writing. This book is the result: I could perhaps have titled it, more whimsically, *A Tale of Independence*, and I would not have been far wrong.

PHILOSOPHY AND THE CUSTOM: The reader who is new to the subject will find a comprehensive exploration of the theory and its rich applications within these covers. But there is something here for the connoisseur as well; any such will find scattered vignettes through the book that will charm, instruct, and illuminate. (This paragraph was written on a day when the sun was shining, the lark was on the wing, and the author was feeling good about the material.)

One of my goals in teaching the subject, and ultimately in writing down what I taught, was to illustrate the intimate connections between abstract theory and vibrant application—there is perhaps no other mathematical science where art, application, and theory coexist so beautifully. This is a serious book withal, an honest book; the proofs of the theorems meet the exacting standards of a professional mathematician. But my pedagogical inclination, shaped by years of teaching, has always been to attempt to discover theorems, perhaps as they might first have been unearthed, rather than to present them one after the other, like so many slices of dry bread, as if they were a mere litany of facts to be noted. And, at the same time, attempt to place the unfolding development of the formal theory in context by both preparing the ground and promptly illustrating its scope with meaty and colourful applications. The novelty is not in new results—although, to be sure, there are new proofs and problems here and there—but in arrangement, presentation, and perspective.

I have endeavoured to make the proofs self-contained and complete, preferably building on repeated elementary themes rather than on a stable of new tricks or sophisticated theorems from other disciplines. It has never struck me that it is pedagogically useful to attempt to prove a theorem by appeal to

another result that the reader has as little chance of knowing; she is being asked to take something on faith in any case and if that is so one may as well just ask her to believe the theorem in question.

Of course there is always the question of what background may be reasonably assumed. My audiences have ranged from undergraduate upperclassmen to beginning graduate students to advanced graduate students and specialists; and they have come from an eclectic welter of disciplines ranging across engineering, computer science, mathematics, statistics, and pure and applied science. A common element in their backgrounds has been a solid foundation in undergraduate mathematics, say, as taught in a standard three or four-course calculus sequence that is a staple in engineering, science, or mathematics curricula. A reader with this as background and an interest in mathematical probability will, with sufficient good will and patience, be able to make her way through most of this book; more advanced tools and techniques are developed where they are needed and a short Appendix fills in lacunae that may have crept into a calculus sequence.

And then there is the question of measure. Probability is, with the possible exception of geometry, the most intuitive of the mathematical sciences and students in a first course on the subject tend to have a strong intuition for it. But, as a graduate student once told me, a subsequent measure-theoretic course on the subject felt as though it were dealing with another subject altogether; a too early focus on measure-theoretic foundations has the unfortunate effect of viewing the subject at a vast remove from its rich intuitive base and the huge application domain, as though measure is from Mars, probability from Venus. And I have found this sentiment echoed repeatedly among students. Something valuable is lost if the price of rigour is a loss of intuition.

I have attempted to satisfy the demands of intuition and rigour in the narrative by beginning with the elementary theory (though a reader should not confuse the word elementary to mean easy or lacking subtlety) and blending in the theory of measure half way through the book. While measure provides the foundation of the modern theory of probability, much of its import, especially in the basic theory, is to provide a guarantee that limiting arguments work seamlessly. The reader willing to take this on faith can plunge into the rich theory and applications in the later chapters in this book, returning to shore up the measure-theoretic details as time and inclination allow. I have found to my pleasant surprise over the years that novices have boldly plunged into passages where students with a little more experience are sadly hampered by the fear of a misstep and tread with caution. Perhaps it is the case that the passage from the novice to the cloister is through confusion.

A serious student of a subject is not an idle spectator to a variety show but learns best by active involvement. This is particularly true of mathematical subjects. I have accordingly included a large collection of problems for solution, scattered quite evenly throughout the text. While there are new problems here

and there to be sure, I have not, by and large, attempted to reinvent the wheel and have taken liberally from the large corpus of problems which are part of the folklore of this venerable subject; providing attribution here is complicated by the confused history and the serial reuse of good problems but where I am aware of a primary source I have provided a name or a reference. Very few of these problems are of a cookbook nature; in some cases they build on the developments in the main body of the text and in others explore significant new areas. Assessing the probable difficulty of a problem for a reader is a tricky task but I have flagged some problems as containing difficult or dangerous digressions so that the reader has some visual guidance.

It has been said with some justification that mathematicians are indifferent historians and, while I cannot claim to have set new standards of accuracy in this regard, I have attempted to provide representative sources; either the original when the antecedents are clear, or a scholarly work which has summarised or clarified the work of many predecessors. Bearing in mind the broad canvas of exploration and the eclectic backgrounds of my students, I have kept the citations generally targeted and specific, not encyclopaedic. While I have faithfully adhered to the original spelling of names in the Latin alphabet, there is no commonly accepted convention for the transliteration of names from other alphabets such as the Cyrillic and variant spellings are to be expected. Bearing in mind Philip J. Davis's admonition in his charming book *The Thread: a Mathematical Yarn* that only admirers of Čaykovskiy's music may write Čebysev in a reference to the great nineteenth-century Russian mathematician, I have kept transliterations phonetic, simple, and common.

Inevitably, the price to be paid for an honest account of the foundational theory and its applications is in coverage. One cannot be all things to all people. I suppose I could plead personal taste in the shape of the narrative but as Kai Lai Chung has remarked in the preface of his classical book on probability, in mathematics, as in music, literature, or cuisine, there is good taste and bad taste; and any author who pleads personal taste must be willing to be judged thereby. And so the discerning reader must decide for herself whether I have been wise in my choices. The reader who wishes to learn more about the theory of Markov chains, renewal theory, information theory, stochastic processes, martingale limit theory, ergodic theory and dynamical systems, or Itô integration must needs look elsewhere. But she will be well prepared with a sound and ample base from which she can sally forth.

The occasional reference to a female reader is idiosyncratic but is not intended to imply that either gender has a monopoly on mathematical thinking; this father was influenced not only by his daughters who over the years peered over his shoulders as he typed and giggled at the titles, but also by the fact that, in life, as in this book, the goddess chance rules.

THE BILL OF FARE; OR, HOW TO READ THIS BOOK: The layout of the text is shown in Figure 1 on the following page. It is arranged in two parts of roughly

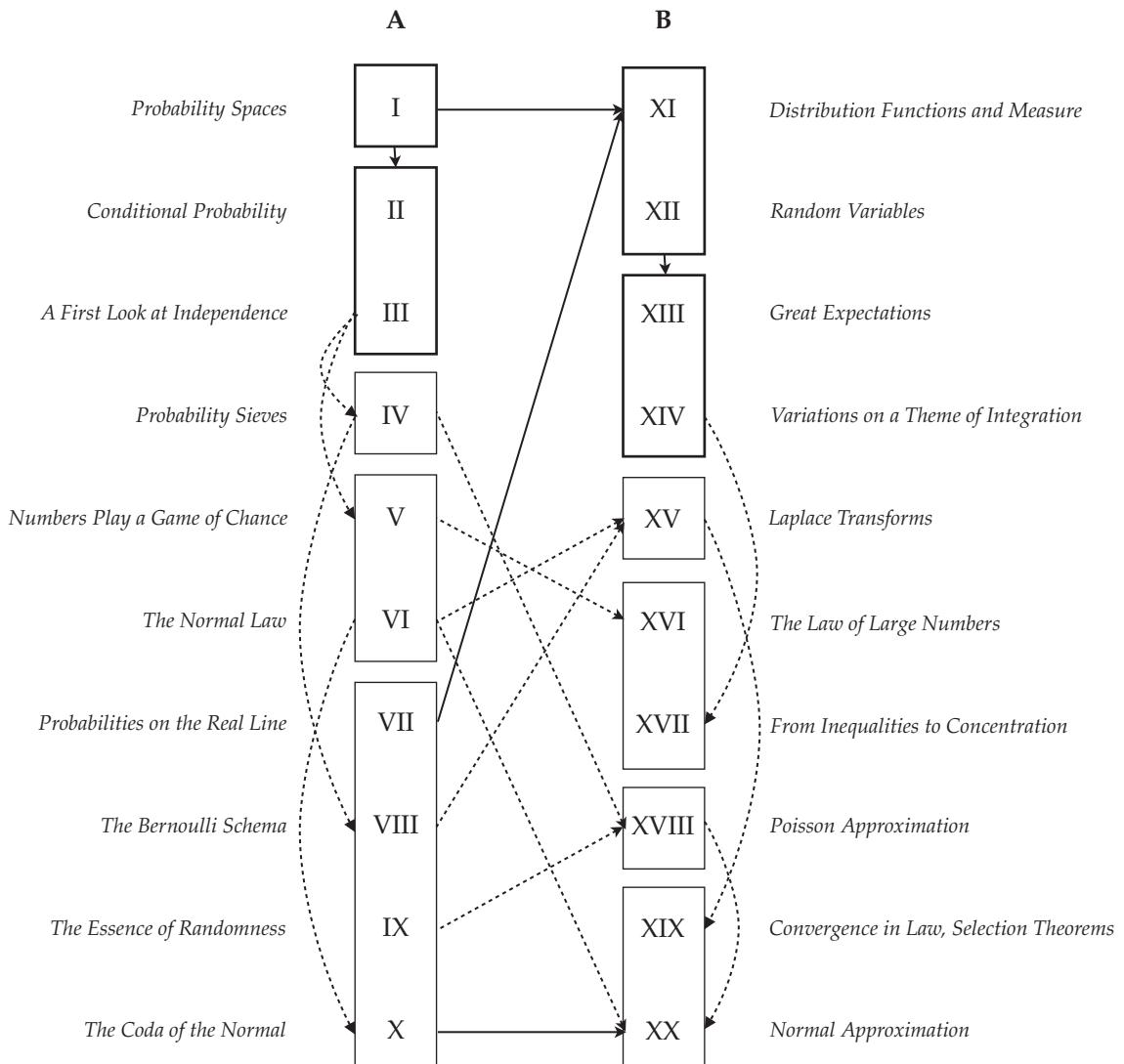


Figure 1: The layout of the book. **Bold** arrows indicate precursor themes which should be absorbed first; dashed arrows indicate connections across themes.

equal size divided into ten chapters apiece. The first part contains the elements of the subject but the more experienced reader will already find previews of deep results obtained by elementary methods scattered through the material: sieve methods and the local lemma (IV); connections with number theory and the laws of large numbers (V); the central limit theorem and large deviations (VI); fluctuation theory (VIII); covering problems and queuing (IX); and mixing and Brownian motion (X). The second part (XI–XX) contains the more abstract foundations where these and other themes are echoed and amplified, and their modern incarnations alluded to earlier fleshed out.

The material was originally written as a sequence of “essays” and while the demands of classroom instruction have meant that the original free-flowing narrative now has some order superimposed upon it in the sequencing of the chapters, the arrangement of the material within and across chapters still has a strong flavour of a menu from which items can be sampled. A reader should study at least the core sections of the introductory chapters of each part (shown enclosed in heavier weight boxes in the figure) before embarking on the succeeding material. With a little good will on the part of the reader each connected block of chapters may then be read independently of the others with, perhaps, a glance at notational conventions of antecedent material. Any such can be readily tracked down via the detailed Index.

To provide further guidance to the reader, the margin of the first page of each chapter contains an annotation cataloguing the nature of the “essays” to follow: \mathcal{C} connotes core sections where key concepts are developed and the basic theorems proved; \mathcal{A} connotes sections containing applications; and, following Bourbaki and Knuth, the “dangerous bend” sigil \diamond connotes sections containing material that is more technical, difficult, or digressive in nature. *A beginning reader should at a minimum read through the core sections;* these contain the “essentials”. Interleaved with these she will find a smorgasbord of applications and digressive material scattered liberally through the text; *these may generally be sampled in any order;* succeeding theory does not, as a general rule, build upon these. To visually guide the reader, section headings for applications and dangerous digressions appear italicised both in page headers and in the Table of Contents. Digressive material is flagged additionally by appearing in small print; *this material can be safely skipped on a first reading without loss of continuity.* Where an entire section is devoted to a tangential or technical tributary, I have also flagged the section with the “dangerous bend” sigil \diamond . Table 1 on the following page shows the breakdown of core, application, and dangerous bend sections.

It is my hope that the more experienced reader will be tempted to flit and explore; for the novice reader there is enough material here for an organised course of study spread over a year if the text is followed linearly. There are also various possibilities for a single-semester course of instruction depending on background and sophistication. I have road tested the following variants.

★ Chapters I–III, VII–X (skipping the bits in small print) can form the core of an

		<i>Chapter Titles</i>	<i>C</i>	<i>A</i>	\ddagger
A	I	<i>Probability Spaces</i>	1–7		8, 9
	II	<i>Conditional Probabilities</i>	1, 3, 7	2, 4–6, 8–10	
	III	<i>A First Look at Independence</i>	1, 4	2, 3	5
	IV	<i>Probability Sieves</i>	1, 4, 6, 9	2, 3, 5, 7, 10	8, 11
	V	<i>Numbers Play a Game of Chance</i>	1–7		
	VI	<i>The Normal Law</i>	1–5		6–9
	VII	<i>Probabilities on the Real Line</i>	1–6, 9, 10	7, 8	
	VIII	<i>The Bernoulli Schema</i>	1, 2, 6, 7, 10	3–5, 9	8
	IX	<i>The Essence of Randomness</i>	1, 5, 8–10	2–4, 6, 7	
	X	<i>The Coda of the Normal</i>	1–4	5, 6, 8, 9	7, 10–12
B	XI	<i>Distribution Functions and Measure</i>	1–3		4, 5
	XII	<i>Random Variables</i>	1–8		9, 10
	XIII	<i>Great Expectations</i>	1–5, 8	6, 7	
	XIV	<i>Variations on a Theme of Integration</i>	1–5, 10	6, 8, 9, 11	7
	XV	<i>Laplace Transforms</i>	1–3	4–8	9
	XVI	<i>The Law of Large Numbers</i>	1, 2, 9	3–7	8, 10–12
	XVII	<i>From Inequalities to Concentration</i>	1, 3–5	2, 7–10	6
	XVIII	<i>Poisson Approximation</i>	1–5, 7	6, 8, 9	
	XIX	<i>Convergence in Law, Selection Theorems</i>	1–3, 5, 6		4, 7–9
	XX	<i>Normal Approximation</i>	1, 5, 8	6, 7, 9–12	2–4
C	XXI	<i>Sequences, Functions, Spaces</i>	1–3		

Table 1: Distribution of sections in the chapter layout. **C:** the core sections of each chapter contain the key concepts, definitions, and the basic theorems; these should be read in sequence. **A:** the application sections are optional and may be sampled in any order. The dangerous bend sections contain subtleties, intriguing, but perilous, examples, technical details not critical to the main flow of the narrative, or simply fun digressions; they should be skipped on a first reading.

honours course for experienced undergraduates, supplemented, at the instructor's discretion, by theory from V and VI or applications from IV, XVI, or XX.

- Chapters I, XI, XII, V, VI, XIII, XIV, and XVI, complemented by selections from X, XV, XIX, or XX, constitute a more abstract, foundational course aimed at graduate students with some prior exposure to probability.
- A seminar course for advanced graduate students can be cobbled together, based on interest, from the following thematic groups of chapters: (XV), (V, XVI, XVII), (IV, XVIII), and (VI, X, XIX, XX).

A word on cross-references and terminology. On pedagogical grounds it seemed to me to be worthwhile to minimise cross-references at the cost of a little repetition. I have accordingly adopted a parsimonious referencing convention and numbered items only where needed for a later reference. Numbered objects like theorems, lemmas, and examples are numbered sequentially *by section* to keep the numbering spare and unencumbered; where needed to unambiguously identify the object under discussion I have amplified the reference to include details of the section or the chapter in which it may be found. To illustrate, Theorem 2 in Section 4 of Chapter IV (this is the first Borel–Cantelli lemma) is referred to in increasing levels of specificity as Theorem 2, Theorem 4.2, or Theorem IV.4.2 depending on whether the reference to it occurs in the same section, another section of the same chapter, or another chapter. Other numbered objects like lemmas, slogans, definitions, examples, and equations are treated in the same way. While I have included a large collection of figures, tables, and problems, these are rarely cross-referenced and I have numbered them sequentially within each chapter; where needed these are identified by chapter.

It is as well to settle a point of terminology here. In keeping with a somewhat cavalier customary usage *I use the terms positive, negative, increasing, and decreasing rather elastically to mean non-negative, non-positive, non-decreasing, and non-increasing, respectively; in these cases I reserve the use of the qualifier "strictly" to eschew the possibility of equality.* Thus, I say that the sequence $\{x_n\}$ is increasing to mean $x_n \leq x_{n+1}$ for each n and modify the statement to $\{x_n\}$ is strictly increasing if I mean that $x_n < x_{n+1}$. Likewise, I say x is positive to mean $x \geq 0$ and say that x is strictly positive when I mean $x > 0$.

ACKNOWLEDGEMENTS: Joel Franklin taught me probability and showed me its beauty. My thinking on the subject was shaped by his gentle guidance, sagacity, and unquenchable enthusiasm. I owe him a deep debt of gratitude.

My students over the years have proved willing foils as I have tested this material on them and my teaching assistants over this period have been pearls beyond price. Shao Fang, Gaurav Kasbekar, Jonathan Nukpezah, Alireza Tahbaz Salehi, Shahin Shahrampour, Evangelos Vergetis, and Zhengwei Wu worked out problems, proofread various portions of the material, and provided suggestions and critiques. Between them they have helped me hone the material and collar ambiguities and errata. Those that remain must, of course, be laid at my door; I can only hope that they are of the venial kind.

My colleagues at Penn and elsewhere have been uniformly encouraging, commenting fulsomely on earlier versions, and I am very grateful for their support and kindness. As has been mentioned by other authors, a long list of acknowledgements excites more attention by its omissions than by its inclusions, and I will not attempt to list everyone to whom I am indebted here. I will only mention that I owe Herb Wilf, Tom Cover, and Mike Steele particular debts of gratitude: Herb, at the inception, for encouraging me to write the book when its outline was barely discernible in a haphazard collection of notes, Tom, at the halfway point when the book was spiralling, apparently ineluctably, into ever-deepening complexity, for reminding me of the virtues of simplicity, and Mike, at the conclusion, for his kind words of encouragement in the dark hour just before dawn when an author looks at his completed work and despairs of it.

I should not forget to mention the good folk at Cambridge University Press who have been so supportive: Phil Meyler waited patiently for this book for a very long time—I can only hope it has been worth the wait—Elizabeth Horne helped select the cover art and provided careful and detailed suggestions on how to improve the look and feel of the manuscript as a whole, Abigail Jones shepherded the book expertly through the production process, and Sarah Lewis whisked the copy-editor’s broom over the document with a light and sure touch.

Part A

ELEMENTS

I

Probability Spaces

Probability is, with the possible exception of Euclidean geometry, the most intuitive of the mathematical sciences. Chance and its language pervades our common experience. We speak of the chances of the weather turning, getting a raise, electing a candidate to political office, or a bill being passed; we bet on sporting contests in office pools, toss a coin at the start of a game to determine sides, wager on the sex of a newborn, and take a chance on the institutionalised gambling that masquerades as state-sponsored lotteries. And, of course, games of chance have an ancient and honoured history. Excavations of bone dice in archaeological digs in North Africa show that dicing was not unknown in ancient Egypt, board games in which players take turns determined by the roll of dice and card games of some antiquity are still popular in the age of the internet, and the horse race survives as an institution. While the historical palette is rich and applications pervasive, the development of a satisfactory mathematical theory of the subject is of relatively recent vintage, dating only to the last century. This theory and the rich applications that it has spawned are the subject of this book.

c 1-7
❧ 8, 9

1 From early beginnings to a model theory

The early history of probability was concerned primarily with the calculation of numerical probabilities for outcomes of games of chance. Perhaps the first book written along these lines was by the eccentric Gerolamo Cardano, a noted gambler, scholar, and bon vivant; his book *Liber de Ludo Aleæ* (Book on Games of Chance) was written perhaps in the 1560s but only published posthumously in 1663.¹ Numerical calculations continued to dominate over the next two and

¹The modern reader will find Cardano's exhortations have weathered well: "The most fundamental principle of all in gambling is simply equal conditions ... of money, of situation ... and of the dice itself. To the extent to which you depart from that equality, if it is in your opponent's favour, you are a fool, and if in your own, you are unjust." This excerpt is from a translation of *Liber de Ludo Aleæ* by Sydney Henry Gould which appears as an appendix in O. Ore, *Cardano, the Gambling Scholar*, Princeton University Press, Princeton, NJ, 1953.

a half centuries awaiting the birth of a theory but the spread of applications continued unabated until, in the modern day, scarce an area of investigation is left untouched by probabilistic considerations.

Today the informed reader encounters probabilistic settings at every turn in divers applications. The following is a representative list of examples, in no particular order, that the reader will find familiar. (i) The conduct of opinion polls—and what the results say about the population as a whole. (ii) Sampling to determine the impact of an invasive species—or of pollutant concentrations. (iii) The prediction of user preferences—for movies or books or soap—from sporadic internet use. (iv) The search for order and predictability in the chaos of financial markets—or of sunspot activity. (v) Robot navigation over uncertain terrain. (vi) The analysis of noise in communications. (vii) The 3K background cosmic radiation and what it portends for the universe. (viii) The statistical physics of radioactive decay. (ix) The description of flocking behaviour in wild geese and fish. (x) The analysis of risk in the design of actuarial tables. (xi) Mendel's theory of heredity. (xii) Genetic combination and recombination, mutation. (xiii) The spread of infection. (xiv) Estimations of time to failure of machinery—or bridges or aeroplanes. (xv) Investment strategies and probabilities of ruin. (xvi) Queues—of telephone calls at an exchange, data packets at an internet server, or cars in a highway system. (xvii) The statistical search for the Higgs boson and other subatomic particles. The reader will be readily able to add to the list from her common experience.

Following the fumbling early beginnings, inevitably numerical, of investigations of the science of chance, as discoveries and applications gathered pace it became more and more necessary that the mathematical foundations of the subject be clearly articulated so that the numerical calculations, especially in areas that could not be readily tested, could be placed on firm mathematical ground. What should the goals of such a theory be? Given the vast realm of applicability we must hold fast against the temptation to hew the theory too close to any particular application. This is much as in how to reach its full flowering geometry had to be sundered from its physical roots. The rôle and importance of abstraction is in the extraction of the logical axiomatic content of the problem divorced of extraneous features that are not relevant, and indeed obfuscate, to provide a general-purpose tool that can be deployed to discover hitherto unsuspected patterns and new directions. Such a clean axiomatic programme was laid out by Andrei Kolmogorov in 1933.²

The key feature of the axiomatic approach is in beginning with a *model* of an idealised *gedanken* chance experiment (that is to say, a thought experiment which may never actually be performed but can be conceived of as being performed). The model, to be sure, makes compromises and introduces convenient mathematical fictions, emphasising certain features of the problem while de-

²English speakers had to wait till 1956 for a translation: A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea, New York.

emphasising or ignoring others, both to permit a clear and unobstructed view of essential features as well as to permit ease of calculation. Thus, for instance, we make the pretence that a coin-tossing game persists indefinitely or that a gambler plays with infinite resources; in the same vein, actuarial tables of lifespans permit aging without bound—albeit with incredibly small probabilities—noise waveforms are modelled as lasting for infinite future time, and so on.

In its insistence on a model for the phenomenon under investigation as a starting point the axiomatic theory makes a clean break with the inchoate idea of intuitive probability that we resort to in our daily experience. The classical wager of Laplace that the sun will rise tomorrow, for instance, has no place in the theory abeyant a reasonable model of a chance experiment (that one can conceive of being in repeated use); similar objections hold for assigning chances to doomsday predictions of, say, terrorist nuclear attacks on major cities, or the destruction of earth by a meteor, or to assigning chances for the discovery of some hitherto unknown propulsion mechanism, and so on. This is unfortunate and we may be reluctant to give up on instinctive, if unformed, ideas of chance in all kinds of circumstances, whether repeatable or not. But as W. Feller has pointed out, “We may fairly lament that intuitive probability is insufficient for scientific purposes but it is a historical fact The appropriate, or ‘natural,’ probability distribution [for particles in statistical physics] seemed perfectly clear to everyone and has been accepted without hesitation by physicists. It turned out, however, that physical particles are not trained in human common sense and the ‘natural’ (or Boltzmann) distribution has to be given up for the [‘unnatural’ or ‘non-intuitive’] Bose–Einstein distribution in some cases, for the Fermi–Dirac distribution in others.”³ The worth of an axiomatic model theory in mathematics is in the rich, unexpected theoretical developments and theorems that flow out of it; and its ultimate worth in application is its observed fit to empirical data and the correctness of its predictions. In these the modern theory of probability has been wildly successful—however unsettling some of its predictions to untrained intuition.

To illustrate the key features of the model it is best to begin with simple chance-driven situations with which we are readily familiar.

2 Chance experiments

Our intuitive assignment of probabilities to results of chance experiments is based on an implicit mathematical idealisation of the notion of repeated independent trials. For instance, in a coin-tossing experiment, conditioned by a complex of experience and custom, we are inclined to treat the coin as “fair”

³W. Feller, *An Introduction to Probability Theory and Its Applications, Volume 1*, 3rd Edition, p. 5. © John Wiley & Sons, 1968. This material is reproduced with permission of John Wiley & Sons, Inc.

and to ascribe probabilities of $1/2$ apiece for heads and tails ignoring possibilities such as that of the coin landing on edge or never coming down at all. Implicit here is the feeling that in a run of n tosses all 2^n possible sequences of heads and tails are equally likely to occur. If in a long run of n tosses there are m heads, we expect that the relative frequency m/n of the occurrence of heads in the tosses will be very close to $1/2$, the accuracy getting better the larger n is.

Now, to be sure, no coin is really “fair”. Statistical analyses of coin flips show invariably that heads and tails are *not* equally likely though the difference tends to be minute in most cases. Nonetheless, the mathematical fiction that a coin is fair is convenient in that it focuses on the essential features of the problem: it is not only simpler analytically but, for most applications, gives predictions that are sufficiently close to reality. We make similar assumptions about the throws of dice in the game of craps, the spin of a roulette wheel, or the distribution of bridge hands in cards. The following simple examples illustrate the key features of the modelling approach.

EXAMPLES: 1) *A coin is tossed three times.* Representing heads by H and tails by T , the possible outcomes of the experiment may be tabulated in a natural convention as HHH , HHT , HTH , HTT , THH , THT , TTH , and TTT . It is clear that these are the only possible outcomes of the idealised experiment and, abeyant any reason to think otherwise, we suppose that all outcomes have equal chance $1/8$ of occurrence. The event that exactly one head is seen may be identified with the aggregate of outcomes consisting of the sequences HHT , THT , and TTH and it is natural to assign to this event the probability $3/8$.

2) *The first throw in craps.* A classical die consists of six faces which we may distinguish by inscribing the numbers 1 through 6 on them (or, as is more usual, by inscribing one through six dots on the faces). The dice game of craps begins by throwing two dice and summing their face values. If the sum of face values is equal to 2, 3, or 12, the player loses immediately; if the sum is 7 or 11, she wins immediately; otherwise, the game continues. What are the chances that a player at craps loses on the first throw? wins on the first throw?

As the only element that decides the result of the first throw is the sum of face values it is natural to consider the outcomes of the experiment (as far as the first throw is concerned) to be the numbers 2 through 12. What are the chances we should ascribe to them? After a little thought the reader may come up with the numbers listed in Table 1. As a loss on the first throw is associated

Outcomes	2	3	4	5	6	7	8	9	10	11	12
Probabilities	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Table 1: The sum of the face values of two dice.

with the aggregate $\{2, 3, 12\}$, it is now reasonable to ascribe to it the probability $\frac{1}{36} + \frac{2}{36} + \frac{1}{36} = \frac{1}{9}$. Similarly, a win on the first throw has associated with it the aggregate of outcomes $\{7, 11\}$ and accordingly has probability $\frac{6}{36} + \frac{2}{36} = \frac{2}{9}$. As any craps player knows, it is twice as likely that she wins on the first throw as that she loses on the first throw.

The critical reader may question the model for the experiment and may prefer instead a model of outcomes as ordered pairs of values, one for each die, the outcomes now ranging over the 36 equally likely possibilities, $(1, 1)$, $(1, 2)$, ..., $(6, 6)$. In this model space, the event of a loss on the first throw may be associated with the aggregate consisting of the pairs $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(6, 6)$, that of a win on the first throw with the aggregate of pairs $(1, 6)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 2)$, $(6, 1)$, $(5, 6)$, and $(6, 5)$. The corresponding probabilities then work out again to be $1/9$ and $2/9$, respectively. A variety of models may describe an underlying chance experiment but, provided they all capture the salient features, they will make the same predictions. All roads lead to Rome. ►

The language of coins, dice, cards, and so on is picturesque and lends colour to the story. But in most cases these problems can be reduced to that of a prosaic placement of balls in urns. The following simple illustration is typical.

EXAMPLE 3) An urn problem. Two balls, say a and b , are distributed in three urns labelled, say, 1, 2, and 3. With the order of occupancy in a given urn irrelevant, the outcomes of the experiment are nine in number, assumed to be equally likely of occurrence, and may be tabulated in the form

$$\begin{aligned} & ab| - | -, \quad -|ab| -, \quad -|-ab, \\ & a|b| -, \quad a|-b, \quad b|a| -, \quad b|-a, \quad -|a|b, \quad -|b|a. \end{aligned} \tag{2.1}$$

The event that the second urn is occupied is described by the aggregate of outcomes $\{-|ab|-, a|b|-, b|a|-, -|a|b, -|b|a\}$ and hence has probability $5/9$. ►

The reader should be able to readily see how the coin and dice problems may be embedded into generic urn problems concerning the placement of n balls into r urns. She may find some fun and profit in figuring out an appropriate urn model for the following catalogue of settings: birthdays, accidents, target shooting, professions (or gender or age), occurrence of mutations, gene distributions, and misprints.

The chance experiments we have considered hitherto deal with a finite number of possible outcomes. But our experience equips us to consider situations with an unbounded number of outcomes as well.

EXAMPLE 4) A coin is tossed until two heads or two tails occur in succession. The outcomes may be tabulated systematically in order of the number of tosses before the experiment terminates, leading to the denumerably infinite list of outcomes in Table 2. If the reader does not immediately believe that the assignment of

<i>Outcomes</i>	HH	TT	THH	HTT	$\text{H}\text{T}\text{H}\text{H}$	$\text{T}\text{H}\text{T}\text{T}$	$\text{T}\text{H}\text{T}\text{H}\text{H}$	$\text{H}\text{T}\text{H}\text{T}\text{T}$	\dots
<i>Probabilities</i>	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$	\dots

Table 2: A sequence of coin tosses terminated when two successive tosses coincide.

probabilities is reasonable, a heuristic justification may be provided by the argument that, if we consider a very long, finite sequence of tosses of length n , a fraction $1/4$ of all such sequences begin with HH and likewise also with TT , a fraction $1/8$ of all such sequences begin with THH and also with HTT , and so on. Allowing n to go to infinity permits the consideration of any terminating sequence of heads and tails. The objection that the experiment cannot in practice last an infinite amount of time so that arbitrarily long sequences are unrealistic in the model may be met with some force by the observation that, for the given probabilities, the chances of requiring more than 100 tosses, say, before termination are $2 \cdot 2^{-100}$. As this will require about 10^{38} performances of the experiment before one such occurrence is detected, one could argue with some justification that most of the assigned probabilities have not been fairly tested. In any case, the reader may well feel that it is even more artificial to fix a stopping point *a priori*, say at 50 tosses, numerical probabilities so chosen as to simply forbid longer sequences by fiat. The practical justification of the model lies in the fact that the assigned probabilities gel nicely with data for sequences of length up to ten or so which carry most of the likelihood; and presumably also for longer sequences though experimental data are abeyant given the fantastically small chances of occurrence.

In this setting, the event that at least four tosses are required before the experiment terminates is captured by the denumerably infinite aggregate of outcomes $\text{H}\text{T}\text{H}\text{H}$, $\text{T}\text{H}\text{T}\text{T}$, $\text{T}\text{H}\text{T}\text{H}\text{H}$, and so on. The probability hence that at least four tosses are required is given by

$$\frac{2}{16} + \frac{2}{32} + \frac{2}{64} + \dots = \frac{2}{16} \left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots \right) = \frac{2}{16} \Big/ \left(1 - \frac{1}{2} \right) = \frac{1}{4},$$

as we identify the infinite sum with a geometric series in powers of $1/2$. Likewise, to the event that the experiment concludes on an odd-numbered trial we may ascribe the probability

$$\frac{2}{8} + \frac{2}{32} + \frac{2}{128} + \dots = \frac{2}{8} \left(1 + \frac{1}{4} + \frac{1}{16} + \dots \right) = \frac{2}{8} \Big/ \left(1 - \frac{1}{4} \right) = \frac{1}{3},$$

as we now encounter a geometric series in powers of $1/4$. It is natural to extend the idea of summing over a finite number of outcome probabilities to a denumerably infinite sum when events are comprised of a countably infinite number of outcomes. ▶

What are the main features we can discern from simple chance experiments of this form? We begin with a model for a *gedanken* experiment whose performance, perhaps only in principle, results in an idealised outcome from a family of possible outcomes. The first element of the model is the specification of an abstract *sample space* representing the collection of idealised outcomes of the thought experiment. Next comes the identification of a family of *events* of interest, each event represented by an aggregate of elements of the sample space. The final element of the model is the specification of a consistent scheme of assignation of *probabilities* to events. We consider these elements in turn.

3 The sample space

R. von Mises introduced the idea of a sample space in 1931⁴ and while his frequency-based ideas of probability did not gain traction—and were soon to be overtaken by Kolmogorov's axiomatisation—the identification of the abstract sample space of a model experiment paved the way for the modern theory.

We shall denote by the uppercase Greek letter Ω an abstract sample space. It represents for us the collection of idealised outcomes of a, perhaps conceptual, chance experiment. The elements ω of Ω will be called *sample points*, each sample point ω identified with an idealised outcome of the underlying *gedanken* experiment. The sample points are the primitives or undefined notions of the abstract setting. They play the same rôle in probability as the abstract concepts of points and lines do in geometry.

The simplest setting for probability experiments arises when the possible outcomes can be enumerated, that is to say, the outcomes are either finite in number or denumerably infinite. In such cases the sample space is said to be *discrete*. The examples of the previous section all deal with discrete spaces.

EXAMPLES: 1) A *coin toss*. The simplest non-trivial chance experiment. The sample space consists of two sample points that we may denote H and T .

2) *Three tosses of a coin*. The sample space corresponding to the experiment of Example 2.1 may be represented by the aggregate HHH, HHT, \dots, TTT of eight sample points.

3) *A throw of a pair of dice*. The sample space consists of the pairs $(1, 1), (1, 2), \dots, (6, 6)$ and has 36 sample points. Alternatively, for the purposes of Example 2.2 we may work with the sample space of 11 elements comprised of the numbers 2 through 12.

4) *Hands at poker, bridge*. A standard pack of cards contains 52 cards in four *suits* (called spades, hearts, diamonds, and clubs), each suit containing 13 distinct

⁴A translation of his treatise was published in 1964: R. von Mises, *Mathematical Theory of Probability and Statistics*, Academic Press, New York.

cards labelled 2 through 10, jack, queen, king, and ace, ordered in increasing rank from low to high. In bridge an ace is high card in a suit; in poker an ace counts either as high (after king) or as low (before 2). A poker hand is a selection of five cards at random from the pack, the sample space consisting of all $\binom{52}{5}$ ways of accomplishing this. A hand at bridge consists of the distribution of the 52 cards to four players, 13 cards per player. From a formal point of view a bridge hand is obtained by randomly partitioning a 52-card pack into four equal groups; the sample space of bridge hands hence consists of $(52!)/(13!)^4$ sample points. In both poker and bridge, the number of hands is so large that repetitions are highly unlikely; the fresh challenge that each game presents contributes no doubt in part to the enduring popularity of these games.

5) *The placement of two balls in three urns.* The sample space corresponding to Example 2.3 may be represented by the aggregate of points (2.1).

6) *The selection of a random graph on three vertices.* A graph on three vertices may be represented visually by three points (or *vertices*) on the plane potentially connected pairwise by lines (or *edges*). There are eight distinct graphs on three vertices—one graph with no edges, three graphs with one edge, three graphs with two edges, and one graph with three edges—each of these graphs constitutes a distinct sample point. A random graph (traditionally represented G_3 instead of ω in this context) is the outcome of a chance experiment which selects one of the eight possible graphs at random. Random graphs are used to model networks in a variety of areas such as telecommunications, transportation, computation, and epidemiology.

7) *The toss of a coin until two successive outcomes are the same.* The sample space is denumerably infinite and is tabulated in Example 2.4. Experiments of this stripe provide natural models for waiting times for phenomena such as the arrival of a customer, the emission of a particle, or an uptick in a stock portfolio. ►

While probabilistic flavour is enhanced by the nature of the application at hand, coins, dice, graphs, cards, and so on, the theory of chance itself is independent of semantics and the specific meaning we attach in a given application to a particular outcome. Thus, for instance, from the formal point of view we could just as well view heads and tails in a coin toss as 1 and 0, respectively, without in any material way affecting the probabilistic statements that result. We may choose hence to focus on the abstract setting of discrete experiments by simply enumerating sample points in any of the standard ways (though tradition compels us to use the standard notation for these spaces instead of Ω).

EXAMPLES: 8) *The natural numbers \mathbb{N} .* The basic denumerably infinite sample space consists of the natural numbers 1, 2, 3,

9) *The integers \mathbb{Z} .* Another denumerably infinite sample space consisting of integer-valued sample points 0, ± 1 , ± 2 ,

10) *The rational numbers \mathbb{Q} .* The sample points are the rational numbers p/q where p is an arbitrary integer and q is a non-zero integer. This space is also denumerable. (If the reader does not know this result she will find it explicated in Example XI.3.2.) ▶

By removing the tether from a particular physical application, an abstract viewpoint permits considerations of broader and richer classes of problems. We may go beyond denumerable spaces of outcomes by considering a limiting sequence of discrete approximations to a continuum of points. In such situations it is simpler to deal directly with the *continuous space* that results. Of course, *gedanken* experiments taking values in a continuum yield much richer sample spaces—and attendant complexities.

EXAMPLES: 11) *The unit interval $[0, 1]$.* When the unit interval is the sample space orthodoxy compels us to call the sample points x (instead of ω); these are the real numbers satisfying $0 \leq x < 1$. Example 7.7 outlines how a natural discrete chance experiment finds a limiting representation in this sample space.

12) *The real line $\mathbb{R} = (-\infty, \infty)$.* The sample points x are now unbounded real numbers $-\infty < x < \infty$. This is a natural extension of the previous example to observables that are modelled as real values.

13) *The complex plane \mathbb{C} .* The sample points are complex numbers $z = x + iy$ where $i = \sqrt{-1}$ and x and y are real. The complex plane is the natural habitat of physical variables such as electric and magnetic fields, currents, and voltages; complex numbers also provide the most economical descriptions of quantum states of matter—a domain where chance is writ large.

14) *Euclidean n-space \mathbb{R}^n .* The sample points $x = (x_1, \dots, x_n)$ are n -tuples of real numbers. This is the natural space when an experiment yields a multiplicity of real observations. ▶

Natural models for a variety of problems lead to even richer sample spaces, for instance, involving functions as the raw objects.

EXAMPLES: 15) *The space \mathcal{C} of continuous functions* whose sample points f are continuous real-valued functions of a real variable. In a variety of problems, one can model random phenomena as producing a *function*, say $f(\cdot)$, as the outcome of an experiment. This is the natural model for noise processes in electrical communications, for instance. The 3K background thermal radiation in the universe that is the observable remnant of the Big Bang may be modelled as a sample function of such a process. The “snow” on a TV screen when a station stops broadcasting is another (two-dimensional) noise process that can be modelled in this fashion.

16) *The space L^2 of square-integrable (or finite-energy) functions* consisting of real (or even complex)-valued functions f with $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$. Physical processes

are necessarily of finite energy and the space L^2 provides a natural model in such settings.

Still richer spaces may be envisaged but with increasing complexity comes increasing difficulty in specifying chance mechanisms. It seems to be a mistake, however, to strive for excessive generality when embarking upon the process of constructing a formal theory and developing an appreciation of its applications. I shall accordingly restrict myself in this book (for the most part) to providing an honest account of the theory and applications of chance experiments in discrete and continuous spaces.

4 Sets and operations on sets

An abstract sample space Ω is an aggregate (or set) of sample points ω . We identify events of interest with subsets of the space at hand. To describe events and their interactions we hence resort to the language and conventions of set theory. We begin with a review of notation and basic concepts.

Accordingly, suppose Ω is an abstract universal set. A subset A of Ω is a subcollection of the elements of Ω . As is usual, we may specify the subsets of Ω by membership, $A = \{\omega : \omega \text{ satisfies a given property } \mathcal{P}\}$, by an explicit listing of elements, $A = \{\omega_1, \omega_2, \dots\}$, or, indirectly, in terms of other subsets via set operations as we detail below. If ω is an element of A we write $\omega \in A$.

We reserve the special symbol \emptyset for the *empty set* containing no elements.

Suppose the sets A and B are subcollections of elements of Ω . We say that A is contained in B (or A is a *subset* of B) if every element of A is contained in B and write $A \subseteq B$ or $B \supseteq A$, both notations coming to the same thing. By convention, the empty set is supposed to be contained in every set. Two sets A and B are *equivalent*, written $A = B$, if, and only if, $A \subseteq B$ and $B \subseteq A$. To verify set equality $A = B$ one must verify both inclusions: first show that any element ω in A must also be in B (thus establishing $A \subseteq B$) and then show that any element ω in B must also be in A (thus establishing $B \subseteq A$). Finally, the sets A and B are *disjoint* if they have no elements in common.

Given sets A and B , new sets may be constructed by disjunctions, conjunctions, and set differences. The *union* of A and B , written $A \cup B$, is the set whose elements are contained in A or in B (or in both). The *intersection* of A and B , written $A \cap B$, is the set whose elements are contained both in A and in B . The *set difference* $A \setminus B$ is the set whose members are those elements of A that are not contained in B ; the special set difference $\Omega \setminus A$ is called the *complement* of A and denoted A^\complement . Finally, the *symmetric difference* between A and B , denoted $A \Delta B$, is the set of points that are contained either in A or in B , but not in both. These operations may be visualised in Venn diagrams as shown in Figure 1.

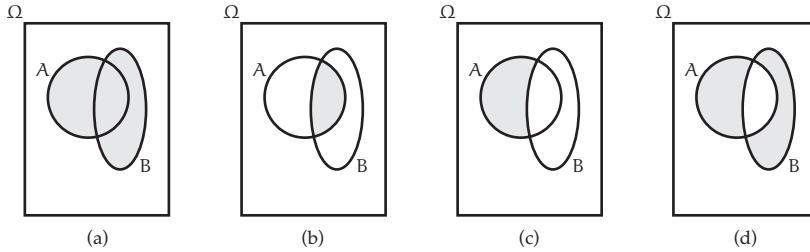


Figure 1: Binary set operations result in the new sets shown shaded. (a) Union: $A \cup B$. (b) Intersection: $A \cap B$. (c) Set difference: $A \setminus B$. (d) Symmetric difference: $A \Delta B$.

Elementary properties of these operations may be set down, almost by inspection. Unions and intersections are commutative, $A \cup B = B \cup A$ and $A \cap B = B \cap A$, associative, $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$, and distribute one over the other, $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

In view of the commutative and associative properties, unions and intersections may be taken in any order and there is accordingly no harm in writing $\bigcup_{k=1}^n A_k$ and $\bigcap_{k=1}^n A_k$ for the union and intersection, respectively, of the sets A_1, \dots, A_n . The notation and intuition extend seamlessly to infinite families of sets. Suppose $\mathcal{A} = \{A_\lambda, \lambda \in \Lambda\}$ is any family of sets indexed by the elements of an index set Λ , possibly infinite. We then write $\bigcup \mathcal{A} = \bigcup_\lambda A_\lambda$ for the set of points ω which belong to at least one set A_λ of the family \mathcal{A} and call it the union of the sets of \mathcal{A} . Likewise, we write $\bigcap \mathcal{A} = \bigcap_\lambda A_\lambda$ for the set of points ω which belong to each of the sets A_λ of the family \mathcal{A} and call it the intersection of the sets of \mathcal{A} . As a matter of convention, if the index set Λ is empty we set $\bigcup_{\lambda \in \emptyset} A_\lambda = \emptyset$ and $\bigcap_{\lambda \in \emptyset} A_\lambda = \Omega$. This somewhat startling convention is motivated by the desire to have the identities

$$\begin{aligned}\bigcup_{\lambda \in \Lambda_1 \cup \Lambda_2} A_\lambda &= \left(\bigcup_{\lambda \in \Lambda_1} A_\lambda \right) \cup \left(\bigcup_{\lambda \in \Lambda_2} A_\lambda \right), \\ \bigcap_{\lambda \in \Lambda_1 \cup \Lambda_2} A_\lambda &= \left(\bigcap_{\lambda \in \Lambda_1} A_\lambda \right) \cap \left(\bigcap_{\lambda \in \Lambda_2} A_\lambda \right),\end{aligned}$$

which are valid whenever the index sets Λ_1 and Λ_2 are non-empty, hold even if one or both of the index sets is empty.

Particularly useful are de Morgan's laws $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$. In tandem with the trite but important observation $(A^c)^c = A$ they allow us to deduce that $A \cap B = (A^c \cup B^c)^c$ so that intersections may be written in terms of unions and complements alone. Likewise, $A \setminus B = (A^c \cup B)^c$ and $A \Delta B = (A^c \cup B)^c \cup (B^c \cup A)^c$ and thus, so can set differences and symmetric differences. The reader who has not been exposed to these identities before should expend a little time in verifying them to settle concepts.

5 The algebra of events

We suppose now that we are dealing with a fixed, abstract sample space Ω representing a conceptual chance experiment.

The *events* of interest are subsets of the space Ω at hand. In the probabilistic setting, the special set Ω connotes the *certain event*, \emptyset the *impossible event*. Probabilistic language lends flavour to the dry terminology of sets. We say that an event A *occurs* if the idealised experiment yields an outcome $\omega \in A$. If A and B are events and $A \subseteq B$ then the occurrence of A implies the occurrence of B . If, on the other hand, $A \cap B = \emptyset$, that is to say, they are disjoint, the occurrence of one precludes the occurrence of the other and we say that A and B are *mutually exclusive*.

Given events A and B , it is natural to construct new events of interest by disjunctions, conjunctions, and set differences. In probabilistic terms: the event $A \cup B$ occurs if A or B or both occur, $A \cap B$ occurs if both A and B occur, $A \setminus B$ occurs if only A occurs (and, in particular, $A^c = \Omega \setminus A$ occurs if A does not), and $A \Delta B$ occurs if precisely one of A and B occur.

Clearly, it is desirable when discussing the family \mathcal{F} of events to include in \mathcal{F} all sets that can be obtained by such natural combinations of events in the family. And as complements and unions suffice to represent any of the other operations, it will be enough for the family \mathcal{F} of events under consideration to be *closed* under complements and unions; that is to say, if A is in \mathcal{F} then so is A^c , and if A and B are in \mathcal{F} then so is $A \cup B$. Any such non-empty family \mathcal{F} of events determines an *algebra* of sets which is closed under the usual set operations in the sense that combinations of events in the natural way lead to other events about which probabilistic statements can be made.

The terminology could use some explication. It is clear that, as long as \mathcal{F} is non-empty, then it must contain both Ω and \emptyset . Indeed, if A is any event in \mathcal{F} then so is A^c whence so is $A \cup A^c = \Omega$ and hence also $(A \cup A^c)^c = \emptyset$. The elements Ω and \emptyset take on the aspect of the unit and the zero in an algebraic system where the operations \cap and Δ play the rôles of multiplication and addition. The interested reader will find the details fleshed out in Problem 23.

Systems of sets closed under unions and complements are also called *fields* in the literature though, for the reasons I have outlined above, from the mathematical purist's point of view they have more in common with algebraic systems.

Considerations of individual events may now be logically extended via induction to finite natural combinations of them. When dealing with infinite sample spaces, however, as we saw in Example 2.4, it is profitable to extend consideration further to *countable* (that is, either finite or denumerably infinite) sequences of such operations. But closure under finite unions does not guarantee closure under countably infinite unions as the following example illustrates.

EXAMPLE 1) *An algebra generated by intervals.* Identifying Ω with the half-closed

unit interval $(0, 1]$, let \mathcal{J} be the set of all half-closed subintervals of the form $(a, b]$. To conform with later usage, let $R(\mathcal{J})$ denote the family of all *finite* unions of the half-closed intervals in \mathcal{J} . The system $R(\mathcal{J})$ is clearly non-empty (and, in particular, contains $(0, 1]$), is manifestly closed under unions, and, as $(a, b]^c = (0, 1] \setminus (a, b] = (0, a] \cup (b, 1]$, it follows that $R(\mathcal{J})$ is closed under complements as well. The system $R(\mathcal{J})$ is hence an algebra. Now let us consider the family of intervals $(0, 1/2]$, $(0, 2/3]$, $(0, 3/4]$, ..., $(0, 1 - 1/n]$, As this is an increasing family, the *finite* union of the first $n - 1$ of these intervals is $(0, 1 - 1/n]$ (and in particular is contained in \mathcal{J} , *a fortiori* also $R(\mathcal{J})$). As we consider these intervals in sequence, any point $0 < x < 1$ contained within the unit interval is eventually contained in an interval $(0, 1 - 1/n]$; indeed, it suffices if $n \geq 1/(1 - x)$. On the other hand, the point $x = 1$ itself is not contained in any of the intervals $(0, 1 - 1/n]$. It follows that the *countably infinite* union $\bigcup_{n=2}^{\infty} (0, 1 - 1/n]$ of all these half-closed intervals is the *open* interval $(0, 1)$ which is *not* contained in $R(\mathcal{J})$. It follows that the algebra $R(\mathcal{J})$ is not closed under countable unions—denumerably infinite set operations on the members of $R(\mathcal{J})$ can yield quite natural sets that are not in it. ▶

Examples of this nature suggest that it would be mildly embarrassing if the family of events that we are willing to consider does not make provision for infinitely extended sequences of events as perfectly legitimate events may then be excluded from the purview of our theory. Bearing in mind this caution, it is prudent to allow ourselves a little wiggle room and entertain denumerably infinite combinations of events. In this context the Greek letter σ (pronounced “sigma”) is used universally to indicate a countable infinity of operations; the origins of the convention may be traced back to the German word *summe*.

DEFINITION A σ -algebra \mathcal{F} is a non-empty family of subsets of Ω , called the *measurable sets* or *events*, satisfying the following two properties:

- If A is a member of \mathcal{F} then so is A^c .
- If $A_1, A_2, \dots, A_n, \dots$ is any countable sequence of sets in \mathcal{F} then their union $\bigcup_{n \geq 1} A_n$ is also in \mathcal{F} .

In other words, a σ -algebra \mathcal{F} contains Ω and is closed under complementation and countable unions. The prefix σ is used here to connote that the algebra of sets is closed even under a denumerably infinite number of operations, not just a finite number. I will leave to the reader the verification that, in conjunction with de Morgan’s laws, a σ -algebra \mathcal{F} is closed under a countable number of any type of binary set operations on its members.

EXAMPLES: 2) *The trivial σ -algebra: $\{\emptyset, \Omega\}$.* Absolutely uninteresting.

3) *The σ -algebra containing a given set.* Let A be any subset of a sample space Ω . Then the family of sets $\{\emptyset, A, A^c, \Omega\}$ forms a σ -algebra as is easy to verify. This is the smallest σ -algebra containing A .

4) *The σ -algebra containing two disjoint sets.* Let A and B be any two disjoint subsets of Ω . The family of sets $\{\emptyset, A, B, A^c, B^c, A \cup B, A^c \cap B^c, \Omega\}$ is the smallest σ -algebra containing A and B . What if $A \cap B \neq \emptyset$?

5) *The total σ -algebra.* The power set of Ω , denoted 2^Ω or $\mathcal{P}(\Omega)$, comprised of all the subsets of Ω (including, of course, the empty set \emptyset and Ω itself) is trivially a σ -algebra. This, the most expansive of all the σ -algebras of the parent set Ω , is called the total σ -algebra. If Ω is a finite set of K elements then the total σ -algebra consists of 2^K subsets. ►

Now it is certainly true that a student can read most of this book assuming that the events at hand are elements of some suitably large σ -algebra of events without worrying about the specifics of what exactly constitutes the family. If she is willing to take the existence of a suitable σ -algebra of events on faith temporarily she should read the following two sections and move on, deferring the study of the material of the final two sections of this chapter till she is ready to embark on the construction of general probability measures in Euclidean spaces in Chapter XII. For the reader who prefers her abstraction served up front, in Section 8 we argue for the necessity of a proper σ -algebra, especially in continuous spaces, and, over the concluding two sections, develop the virtues of a more abstract point of view and its consequences.

6 The probability measure

So we now begin with an abstract sample space Ω equipped with a σ -algebra \mathcal{F} containing the events of interest to us. The final step in Kolmogorov's programme is to determine a consistent scheme of assigning probabilities to events. As always, it is prudent not to assume any more structure than intuition guided by experience suggests is necessary.

A *probability measure* is a map $A \mapsto \mathbf{P}(A)$ which, to each event $A \in \mathcal{F}$, assigns a value $\mathbf{P}(A)$ (in other words, $\mathbf{P}: \mathcal{F} \rightarrow \mathbb{R}$ is a *set function*) satisfying the following axioms.

Axiom 1 Positivity: $\mathbf{P}(A) \geq 0$ for every event $A \in \mathcal{F}$.

Axiom 2 Normalisation: $\mathbf{P}(\Omega) = 1$.

Axiom 3 Additivity: Suppose A and B are disjoint events, $A \cap B = \emptyset$. Then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$.

Axiom 4 Continuity: Suppose A_1, A_2, \dots is a sequence of events decreasing monotonically to the empty set; that is to say, $A_n \supseteq A_{n+1}$ for each n and $\bigcap_{n \geq 1} A_n = \emptyset$. Then $\mathbf{P}(A_n) \rightarrow 0$ as $n \rightarrow \infty$.

The triple $(\Omega, \mathcal{F}, \mathbf{P})$ is called a *probability space* though, from a logical point of view, the specification is redundant: the σ -algebra \mathcal{F} implicitly determines the underlying sample space so that it is not necessary to explicitly specify Ω , while the probability measure \mathbf{P} identifies the particular σ -algebra \mathcal{F} under consideration (and thence also the sample space Ω) so that it is not, strictly speaking, necessary to specify \mathcal{F} explicitly either—the probability measure \mathbf{P} implicitly identifies all three elements of the probability space. As a reaction against excessive formalism, it does no harm if we take a liberty and say that Ω is a probability space (by which we mean, of course, that we are working with a particular σ -algebra \mathcal{F} on which a probability measure \mathbf{P} is defined) and, in the same vein, that \mathbf{P} is a probability measure on Ω (implicitly, of course, on a particular σ -algebra \mathcal{F} of Ω).

How do these axioms gel with our experience? Our intuitive assignment of probabilities to results of chance experiments is based on an implicit mathematical idealisation of the notion of repeated independent trials. Suppose A is an event. If, in n independent trials (we use the word “independent” here in the sense that we attach to it in ordinary language) A occurs m times then it is natural to think of the relative frequency m/n of the occurrence of A as a measure of its probability. Indeed, we anticipate that in a long run of trials the relative frequency of occurrence becomes a better and better fit to the “true” underlying probability of A .

There are clearly difficulties with this “frequentist” approach to probability, not the least of which is that two different sequences of trials may give different relative frequencies; nor is there any absolute guarantee that the sequence of relative frequencies will actually converge in the limit of a large number of trials. The ultimate justification of this intuitive point of view will be provided by the law of large numbers after we have developed some background but certain features are already suggestive.

As $0 \leq m/n \leq 1$, the positivity and normalisation axioms are natural if our intuition for odds in games of chance is to mean anything at all. The selection of 1 as normalisation constant is a matter of convention. While other normalisation constants are possible—in casual conversation we use a baseline of 100 when we refer to odds as percentages—in probability theory the choice of 1 as constant is universal. Likewise, from the vantage point of relative frequencies, if A and B are two mutually exclusive events and if in n independent trials A occurs m_1 times and B occurs m_2 times, then the relative frequency of occurrence of either A or B in n trials is $\frac{m_1+m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n}$ and stands as a surrogate for the probability of the disjunction $A \cup B$. Probability measure is now forced to be additive if it is to be consistent with experience. And, finally, the continuity axiom is so natural as to not evoke comment; after all, it is entirely reasonable to suppose that the smaller the relative frequency of an event, the rarer its chance of occurrence, the impossible event occurring never. The continuity axiom is so natural indeed that the reader may wonder whether it

is implicit in the others. In actual fact it is logically independent of the other axioms; this is explored further in the *Problems*.

Before proceeding it will be useful to take stock of some immediate implications of the axioms.

To begin, while the axioms do not say anything explicit about the empty set, as $\emptyset = \emptyset \cup \emptyset = \emptyset \cap \emptyset$, we may deduce by additivity of probability measure that $P(\emptyset) = P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset)$, which relation implies immediately that *the empty set has probability zero*: $P(\emptyset) = 0$. Satisfactory.

Now suppose that the occurrence of an event A implies the occurrence of B , that is to say, $A \subseteq B$. Then B may be written as the disjoint union of A and $B \setminus A$ whence $P(B) = P(A) + P(B \setminus A) \geq P(A)$ by additivity and positivity. It follows that probability measure indeed has the intuitive and important *monotone property*: *if A and B are events in a probability space and $A \subseteq B$ then $P(A) \leq P(B)$* .

The continuity axiom enables us to extend additivity from finite collections of disjoint sets to countably infinite disjoint collections.

THEOREM *If $\{A_n, n \geq 1\}$ is a sequence of pairwise disjoint events in some probability space then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.*

PROOF: Let $B_1 = A_1$ and, for each $n \geq 2$, let $B_n = B_{n-1} \cup A_n = A_1 \cup \dots \cup A_n$. The events B_{n-1} and A_n are disjoint so that, by repeated application of additivity of probability measure, we have

$$P(B_n) = P(B_{n-1}) + P(A_n) = P(A_1) + \dots + P(A_{n-1}) + P(A_n).$$

Passing to the limit as $n \rightarrow \infty$, we hence obtain

$$\lim_{n \rightarrow \infty} P(B_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n P(A_k) = \sum_{k=1}^{\infty} P(A_k),$$

the convergence of the series on the right guaranteed because, by positivity, $P(B_1) \leq P(B_2) \leq \dots \leq P(B_n) \leq \dots \leq 1$ so that $\{P(B_n), n \geq 1\}$ is a bounded, increasing sequence, hence has a limit.

Setting $A = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$, we now consider the events $C_n = A \setminus B_n = \bigcup_{k=n+1}^{\infty} A_k$. Then C_1, C_2, \dots is a sequence of events decreasing monotonically to the empty set. (Why? Any sample point $\omega \in A_n$ does not lie in C_{n+1}, C_{n+2}, \dots and hence each sample point can lie in only a finite number of the C_n .) By continuity of probability measure it follows that $\lim_{n \rightarrow \infty} P(C_n) = 0$. As A may be written as the disjoint union of B_n and C_n , additivity shows that $P(A) = P(B_n) + P(C_n)$. Taking limits of both sides completes the proof. ▶

The abstract point of view embodied in the four axioms certainly eliminates the ambiguities inherent in our intuitive idea of chance. Its ultimate worth, however, depends on whether the model yields useful predictions in real applications and whether it yields fresh insights into and a deeper understanding of the mechanism of chance. This is the subject of this book.

7 Probabilities in simple cases

Viewed as a function *qua* function, probability measure is undeniably a very complicated object. How should one in general proceed to construct a set function consistent with the four axioms? I will begin by considering two simple but important cases to help build intuition.

DISCRETE SPACES

The simplest situation arises when the sample space Ω is either finite or at most denumerably infinite. In this case we may represent Ω by a set of idealised outcomes ω_k where k ranges over a finite or denumerably infinite set of values and we say that Ω is *countable* or *discrete*. There are no technical complications here in allowing the σ -algebra to consist of all subsets of Ω and we may as well do so. The construction of a probability measure is particularly transparent in this setting.

Suppose then that $\Omega = \{\omega_k, k \in \mathbb{Z}\}$ is a countable sample space. Let $\{p_k, k \in \mathbb{Z}\}$ be any summable sequence of positive numbers, $p_k \geq 0$, so normalised that $\sum_k p_k = 1$, the sum running over all integers k . With each such sequence we may associate the map $A \mapsto P(A)$ which, to each subset A of $\Omega = \{\omega_k, k \in \mathbb{Z}\}$, assigns the value

$$P(A) = \sum_{k: \omega_k \in A} p_k$$

(where the sum is over all indices k for which ω_k is in A). If p_k is non-zero only for a finite number of values k the situation reduces to that of a finite sample space. It is easy to see that P is a probability measure on the subsets of Ω .

Indeed, the positivity, normalisation, and additivity axioms are trivially satisfied by the constructed set function P . It only remains to verify the continuity axiom. Suppose $\epsilon > 0$ is a prototypical, small positive number. As the sequence $\{p_k, k \in \mathbb{Z}\}$ is summable, for a sufficiently large value of $N = N(\epsilon)$, we have $\sum_{k:|k| \geq N+1} p_k < \epsilon$. Now suppose that $\{A_n, n \geq 1\}$ is a decreasing sequence of subsets of $\{\omega_k, k \in \mathbb{Z}\}$ with $\bigcap_n A_n = \emptyset$. It is clear that each of the $2N+1$ sample points $\omega_{-N}, \omega_{-N+1}, \dots, \omega_{N-1}, \omega_N$ can lie in at most a finite number of the sets A_n . Accordingly, if n is sufficiently large, none of the sample points $\omega_{-N}, \dots, \omega_N$ will lie in A_n and *a fortiori* in any of the sets A_{n+1}, A_{n+2}, \dots thereafter. It follows that $A_n \subseteq \{\omega_k : |k| \geq N+1\}$ for all sufficiently large n and, as $\{p_k, k \in \mathbb{Z}\}$ is a sequence of positive numbers,

$$P(A_n) = \sum_{k: \omega_k \in A_n} p_k \leq \sum_{k: |k| \geq N+1} p_k < \epsilon.$$

As ϵ may be taken arbitrarily small, this means that $P(A_n) \rightarrow 0$ as $n \rightarrow \infty$ and it follows that the set function P is continuous. The examples of Section 2 all deal with discrete probability distributions.

Each positive, summable, and normalised sequence $\{p_k, k \in \mathbb{Z}\}$ hence determines a probability measure on the discrete space Ω . The constructed measure P maps each single-element set $\{\omega_k\}$ into the value $P\{\omega_k\} = p_k$.⁵ Informally we say that the outcome ω_k has probability p_k . The sequence $\{p_k, k \geq 1\}$ is hence called a *discrete probability distribution*. In this light, the map $A \mapsto P(A)$ associates with A the sum of probabilities of those outcomes that comprise A .

For any discrete space $\Omega = \{\omega_k, k \in \mathbb{Z}\}$ the specific nature of the sample points is irrelevant as far as the probability distribution $\{p_k, k \in \mathbb{Z}\}$ is concerned and, for concreteness, we identify the discrete space Ω with the family of integers \mathbb{Z} in the examples that follow.

EXAMPLES: 1) *Combinatorial probabilities.* If $p_0 = p_1 = \dots = p_{n-1} = 1/n$ (all other p_k being identically zero) then the sample space is finite and consists of n equally likely outcomes $0, 1, \dots, n - 1$. The probability of any event A is then given by $\text{card}(A)/n$ (where $\text{card}(A)$ is the *cardinality* or size of A) and the computation of probabilities devolves into a counting exercise. This is the classical setting of combinatorial probabilities.

2) *Bernoulli trials.* If $p_0 = 1 - p$ and $p_1 = p$ where $0 \leq p \leq 1$, the sample space consists of two elements 0 and 1 (or tails and heads). This is the setting of the toss of a bent coin whose success probability is p .

3) *The binomial distribution.* Again with $0 \leq p \leq 1$, the choice $p_0 = (1-p)^n, p_1 = np(1-p)^{n-1}, \dots, p_k = \binom{n}{k}p^k(1-p)^{n-k}, \dots, p_n = p^n$ specifies a probability distribution on a finite set of $n + 1$ elements, $0, 1, \dots, n$. This is the binomial distribution with p_k representing the chances of obtaining exactly k heads in n tosses of the bent coin of the previous example.

4) *Waiting time distributions.* The distribution obtained by setting $p_k = (1-p)^k p$ for $k \geq 0$ represents the distribution of the number of failures before the first success in a sequence of tosses of the coin of Example 2. This is the *geometric distribution*. More generally, for any integer $r \geq 1$, the distribution defined by setting $p_k = \binom{r+k-1}{k}(1-p)^k p^r$ for $k \geq 0$ represents the distribution of the number of failures before the r th success in a sequence of tosses of the coin. This is the *negative binomial distribution*.

5) *The Poisson distribution.* Suppose $\lambda > 0$. If, for each positive integer $k \geq 0$, we set $p_k = e^{-\lambda} \lambda^k / k!$, we obtain the Poisson distribution on the positive integers. The Poisson distribution crops up in the characterisation of rare events. ►

We shall revisit these distributions in somewhat more depth in Chapter VIII. The next example returns to the theme of a discrete sequence of coin tosses—with an unexpected twist.

⁵Formally, $P(\{\omega_k\})$ is the probability of the singleton $\{\omega_k\}$ but such pedantry rapidly gets wearing. There is no risk of confusion in omitting the round parentheses in the argument of P when sets are explicitly specified using curly brackets and we do so without further ado.

EXAMPLES: 6) *Coin tosses, an unexpected quandary.* Consider a *gedanken* experiment corresponding to an unending sequence of tosses of a fair coin. The sample space Ω is the collection of all unending sequences of the form $z_1 z_2 z_3 \dots$ where each z_n is either heads or tails. As before, we associate any event with a subset of the sample space, i.e., a collection of sequences sharing the property of interest. Thus, for instance, if A_4 is the event that the first four tosses of the sequence result in tails then on intuitive grounds one is tempted to assign a probability of $1/16$ to A_4 . More generally, let A_n be the event that the first n tosses of the sequence result in tails. On the same intuitive grounds, for each n , it is natural to assign to A_n the probability 2^{-n} .

What exactly are the sets A_n ? As the sample points of the conceptual chance experiment are unending sequences of heads and tails, the event A_n consists precisely of those sequences $z_1 z_2 \dots z_n z_{n+1} \dots$ where, with \mathfrak{H} denoting heads and \mathfrak{T} denoting tails, $z_1 = \dots = z_n = \mathfrak{T}$ with the values $z_k \in \{\mathfrak{H}, \mathfrak{T}\}$ for $k \geq n + 1$ arbitrarily specified. Now it is clear that the events A_1, A_2, \dots form a decreasing sequence of sets, $A_n \supseteq A_{n+1}$, where the limiting set $\bigcap_n A_n$ is the single-element set consisting of that sequence all of whose components are tails. By continuity of probability measure (or, more precisely, Problem 29), it follows that the probability of an unending sequence of tails is given by

$$P\{\mathfrak{T}\mathfrak{T}\mathfrak{T}\mathfrak{T}\dots\} = P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} 2^{-n} = 0.$$

The same argument works for any sequence of toss outcomes: if $z_1^0 z_2^0 z_3^0 \dots$ is any sequence of heads and tails, we may define the event A_n as the collection of sequences $z_1 \dots z_n z_{n+1} \dots$ for which the first n tosses are precisely $z_1 = z_1^0, \dots, z_n = z_n^0$, with all subsequent values z_{n+1}, z_{n+2}, \dots arbitrarily specified. Again, we would like to associate the probability 2^{-n} with A_n and the continuity axiom shows as before that the single-element probability $P\{z_1^0 z_2^0 z_3^0 \dots\}$ is zero. We are then led inexorably to the conclusion that *all sample points (i.e., unending sequences of heads and tails) have probability identically zero*. If we now attempt to write down the probability of, say, the event A_4 as the sum of the probabilities of the individual sequences comprising it we seem to be led to the paradoxical result that $1/16 = 0$. How should we read this riddle?

7) *Coin tosses, resolution.* As is frequently the case, a change of perspective makes the situation transparent. The reader will recall that any real number $0 \leq t < 1$ may be represented in a dyadic expansion in base 2 of the form

$$t = z_1 2^{-1} + z_2 2^{-2} + \dots + z_n 2^{-n} + \dots = .z_1 z_2 \dots z_n \dots \quad (\text{base 2})$$

where each z_n takes a value 0 or 1 only. (There is a minor nuisance in that points of the form $m/2^n$ have two representations; for instance, $3/8$ may be written in either of the forms $.011\dot{0}$ or $.010\dot{1}$, the dot indicating that the number

repeats indefinitely. In such cases, for definiteness, we select the former representation with 0 repeating indefinitely.) If we rename heads and tails by 1 and 0, respectively, then each real number t in the unit interval may be mapped into an outcome $z = z_1 z_2 \dots$ of our conceptual experiment. The sample points of our experiment hence span the continuum of points in the unit interval and, quite unexpectedly, we discover that *the sample space of an unending sequence of coin tosses is continuous!* Events of natural interest in this setting may be identified with intervals; thus, for instance, the event A_4 that the first four tosses are tails may be identified with the set of points t satisfying $.0000\bar{0} \leq t < .0001\bar{0}$ or, equivalently, the interval $[0, 1/16)$ (bearing in mind the convention that of the two representations $1/16 = .0001\bar{0} = .0000\bar{1}$, we select the former); the event A_n , likewise, may be identified with the interval $[0, 2^{-n})$. More generally, the events of interest correspond to finite, or even countable, collections of subintervals of the unit interval.

With the identification of the sample space with the unit interval $[0, 1)$ and events with subintervals, the natural probability measure is to associate with each event the length of the interval corresponding to it. It is easy to see that, at least when restricted to considerations of intervals, length brings with it the accustomed features of positivity, additivity, and continuity that we associate with probability measure. This suffices for our immediate purposes; a more searching examination of what length means in general will have to wait for developments in Chapter XII.

According to our prescription, to the event A_4 we may associate as probability the length of the interval $[0, 1/16)$; the procedure certainly seems strange given the very discrete coin-tossing flavour of the experiment, but our confidence is bolstered by the fact that the value $1/16$ that results is quite in accordance with our coin-tossing intuition! To take another example, the event A that the first head occurs on the fourth toss may be identified with the set of points t for which $.0001\bar{0} \leq t < .001\bar{0}$ or, equivalently, the interval $[1/16, 1/8)$. The probability of A is hence given by $P(A) = 1/8 - 1/16 = 1/16$ as is again natural and intuitive.

This then is the resolution of our riddle: the event A_4 , for instance, may be mapped onto the continuum interval $[0, 1/16)$ and it follows that the elements comprising A_4 are *not* countable. As the reader is well aware, when one moves to the continuum, the addition or deletion of individual points t does not affect interval length—the points t of the interval $[0, 1/16)$ individually have zero length but, viewed en masse over the uncountable continuum of points comprising the interval, the length of the interval is strictly positive. ►

This seemingly accidental conjunction of coin-tossing experiments and number systems has deep consequences. We will explore the subject further in Chapters V and VI.

Our example shows that very rich sample spaces spanning a continuum of outcomes can arise out of experiments with impeccable discrete proveances. In such situations it is simpler to deal directly with the continuous space instead of a clumsy sequence of discrete approximations.

CONTINUOUS SPACES

If the outcomes of a conceptual probability experiment take a continuum of values on the real line we shall refer to the space as *continuous* and identify the sample space Ω with the real line $\mathbb{R} = (-\infty, \infty)$. In this setting our focus on individual sample points gives way naturally to intervals which now become the basic objects of interest, events then connoting special subsets of the line. Point probabilities now give way to probability *densities* in units of probability mass per unit length, discrete sums segueing in consequence into integrals.

As a matter of notation, we visually mark the special and important character of subsets of the line \mathbb{R} and, more generally, Euclidean space \mathbb{R}^n , by using Blackboard bold fonts such as \mathbb{A} , \mathbb{B} , \mathbb{I} , and \mathbb{J} to identify them.

Let f be any positive and integrable function defined on the real line \mathbb{R} and normalised so that $\int_{-\infty}^{\infty} f(x) dx = 1$. With each such function we may associate the probability map $\mathbb{A} \mapsto P(\mathbb{A})$ which, to each subset \mathbb{A} of the line of interest, assigns the value

$$P(\mathbb{A}) = \int_{\mathbb{A}} f(x) dx.$$

For the time being we shall interpret integrals of the above form as common or garden variety Riemann integrals. There is surely no difficulty in the interpretation of the integral if f is piecewise continuous and \mathbb{A} is simply an interval or a finite (or even countable) union of disjoint intervals and we will restrict ourselves to such situations for the nonce. We will eventually need a more flexible and powerful notion of integration that will enable us to handle integration over general subsets \mathbb{A} of the real line (the Borel sets) as well as a richer class of continuous distributions and we will turn to a flexible and powerful notion of integration pioneered by Henri Lebesgue for the purpose.

Thus, if \mathbb{I} is any of the four types of intervals (a, b) , $[a, b)$, $[a, b]$, or $(a, b]$, we obtain $P(\mathbb{I}) = \int_{\mathbb{I}} f(x) dx = \int_a^b f(x) dx$, the integral being insensitive to the addition or deletion of points on the boundary, while, if $\{\mathbb{I}_n, n \geq 1\}$ is a countable collection of pairwise disjoint intervals then $P(\bigcup_n \mathbb{I}_n) = \sum_n \int_{\mathbb{I}_n} f(x) dx$ as integration is additive. It follows that the constructed set function P is indeed positive, additive, and continuous, at least when restricted to intervals and finite or countable unions of them. We will take it on faith temporarily that these familiar properties of integration are preserved even for integrals over general (Borel) sets on the line (however such integrals are to be interpreted).

When the sample space is continuous, the positive, integrable, and normalised function f assumes the rôle played by the discrete probability distribution $\{p_k, k \in \mathbb{Z}\}$ in the discrete case. The value $f(x)$ assumed by f at a given

point x of the line connotes the probability mass per unit length that is attached at that point; in Leibniz's language of infinitesimals $f(x) dx$ represents the probability that an outcome of the conceptual continuous experiment takes a value in an infinitesimal interval of length dx at the point x . The function f is hence called a *probability density* (or simply *density* in short). The intuition very much parallels what one sees in physics when one progresses from point masses to a continuum of mass. In moving from the discrete to the continuum, sums of probability masses segue naturally into integrals of the density.

EXAMPLES: 8) *The uniform density.* The function $f(x)$ which takes value $1/(b-a)$ if $a < x < b$ and value 0 otherwise attaches a uniform mass density to the interval (a, b) . It may be taken, following Laplace, as a statement of agnosticism about the outcome of a chance experiment with values in the interval (a, b) . The density implicit in Examples 6 and 7 is uniform in the unit interval.

9) *The exponential density.* If $\alpha > 0$, the one-sided exponential function taking value $f(x) = \alpha e^{-\alpha x}$ for $x > 0$ and value 0 otherwise plays an important rôle as providing a model for true randomness.

10) *The normal density.* The function $f(x) = (2\pi)^{-1/2} \sigma^{-1} e^{-(x-\mu)^2/2\sigma^2}$ (where μ is real and σ is strictly positive) is of central importance. That it is integrable is easily seen from the rapid decay of the exponential $e^{-x^2/2}$ away from the origin. The demonstration that it is properly normalised to form a density relies upon a simple trick which we shall defer to Section VI.1. ▶

We will return to a more searching examination of these densities and the contexts in which they appear in Chapters IX and X.

The notation and language carry over smoothly to continuous sample spaces in any number of dimensions. Thus, if we identify the sample space Ω with the n -dimensional Euclidean space \mathbb{R}^n the density f represents a positive, integrable, and normalised function of n variables and all that is useful is to interpret all integrals as integrals over regions in \mathbb{R}^n .



8 Generated σ -algebras, Borel sets

Given a sample space Ω , why not simply deal with the total σ -algebra $\mathcal{P}(\Omega)$ of all possible subsets? Indeed, for discrete spaces we may just as well consider the totality of all subsets as events; the assignment of probabilities is simple in this case as we have seen. For continuous spaces, however, the family $\mathcal{P}(\Omega)$ is too rich to handle for it transpires that it proves impossible to assign probabilities in a consistent fashion to all possible subsets. How should one then proceed? In order to construct more interesting event families which are useful and yet are not so rich as to be intractable, one proceeds indirectly by starting with a family of sets of natural interest dictated by the application at hand. In accordance with the fourteenth-century recommendations of William of Ockham, we adopt parsimony as a guide and propose to use the *smallest* σ -algebra that will serve.

Accordingly, we begin with any family \mathcal{A} of subsets of Ω of interest. Our objective is now to identify a minimal σ -algebra which contains all the sets of \mathcal{A} among its elements. We proceed indirectly. Let $\mathcal{S}(\mathcal{A})$ denote the family of σ -algebras \mathcal{F} each of which includes all the sets of \mathcal{A} . The family $\mathcal{S}(\mathcal{A})$ is not empty as it certainly contains the total σ -algebra of all subsets of Ω as one of its members. Now let $\sigma(\mathcal{A})$ denote the collection of subsets A of Ω which are contained in every $\mathcal{F} \in \mathcal{S}(\mathcal{A})$, that is to say, $\sigma(\mathcal{A}) = \bigcap_{\mathcal{F} \in \mathcal{S}(\mathcal{A})} \mathcal{F}$ is the intersection of the sets \mathcal{F} in the family $\mathcal{S}(\mathcal{A})$. (The reader will bear in mind that each element \mathcal{F} of the family $\mathcal{S}(\mathcal{A})$ is itself a set.) As the sets in \mathcal{A} are contained in each of the σ -algebras \mathcal{F} in $\mathcal{S}(\mathcal{A})$, it is clear that $\sigma(\mathcal{A})$ is non-empty and, at a minimum, contains all the sets of \mathcal{A} . We now argue that $\sigma(\mathcal{A})$ is itself a σ -algebra. Indeed, if A is in $\sigma(\mathcal{A})$ then it must *a fortiori* lie in every \mathcal{F} in $\mathcal{S}(\mathcal{A})$; and, as each \mathcal{F} is itself a σ -algebra and hence closed under complementation, it must also contain A^c . Consequently, every set $\mathcal{F} \in \mathcal{S}(\mathcal{A})$ contains both A and A^c ; and hence so does $\sigma(\mathcal{A})$. Now if $\{A_n, n \geq 1\}$ is any countable sequence of sets in $\sigma(\mathcal{A})$ then every $\mathcal{F} \in \mathcal{S}(\mathcal{A})$ contains $\{A_n\}$, hence contains $\bigcup_n A_n$, whence $\sigma(\mathcal{A})$ also contains $\bigcup_n A_n$. Thus, $\sigma(\mathcal{A})$ is non-empty and closed under complementation and countable unions, or, in other words, it is a σ -algebra. As $\sigma(\mathcal{A})$ is a subset of every σ -algebra \mathcal{F} in the family $\mathcal{S}(\mathcal{A})$, it is the *smallest* σ -algebra containing \mathcal{A} ; we call $\sigma(\mathcal{A})$ the σ -algebra *generated by* \mathcal{A} .

THEOREM Suppose \mathcal{A} is any non-empty family of sets. Then $\sigma(\mathcal{A})$ is the unique minimal σ -algebra containing all the sets in \mathcal{A} .

EXAMPLE 1 The σ -algebra of Borel sets. Intervals and combinations of intervals carry a natural importance in the setting of a real-line sample space, $\Omega = \mathbb{R}$, and any reasonable event family should surely include them. Accordingly, reusing notation from Example 5.1, consider the set \mathcal{I} of half-closed intervals of the form $(a, b]$ where, now, $-\infty < a < b < \infty$. The natural family $\mathcal{R}(\mathcal{I})$ of sets obtained by expanding \mathcal{I} to include all finite unions of half-closed intervals is manifestly closed under finite unions and (as $(a, b] \setminus (c, d]$ is either empty or another half-closed interval) also closed under set differences but is not quite sufficient for our purposes as it is not closed under complementation or countable unions and does not include \mathbb{R} among its elements. In view of our theorem, however, we can expand \mathcal{I} further until we have a bona fide minimal σ -algebra $\mathcal{B} = \sigma(\mathcal{I})$ containing the half-closed intervals. This is called the *Borel σ -algebra* and its elements are called the *Borel sets (of the line)*. ▶

Three questions may have crossed the reader's mind at this juncture.

Why the focus on the half-closed intervals $(a, b]$? Well, any other kind of interval, open (a, b) or closed $[a, b]$, would do equally well for our purposes. Indeed, any one of the interval types can be systematically deployed to generate the others; see Problem 26. However, as remarked earlier, the family \mathcal{I} of half-closed intervals has the desirable property that it is closed under set differences, a feature that does much to alleviate purely technical irritations. This would not be true if we had chosen to populate \mathcal{I} with either open or closed intervals. The choice of the left-open, right-closed intervals $(a, b]$ over the left-closed, right-open intervals $[a, b)$ is merely traditional. Either would serve.

Are all Borel sets representable by a countable sequence of operations on intervals? The Borel σ -algebra \mathcal{B} generated by the family of half-closed intervals \mathcal{I} is a very complex beast. It contains, among other things, all the other kinds of intervals, open, closed, and (left) half-closed, and also all singleton sets consisting of individual points on the line;

see Problem 26. And, of course, it also contains countable unions, intersections, set differences, and symmetric differences of all these objects. However, it is a sad fact, but true, that the requirements of closure under complementation and countable unions force \mathcal{B} to contain sets much wilder than these: *not all Borel sets can be obtained by a countable sequence of set operations on intervals.* In most applications, however, the natural events of interest are directly obtainable by a countable number of operations on intervals and the more esoteric Borel sets need not concern us.

The Borel σ -algebra appears very complex: does it, by any chance, include all the subsets of the real line? In other words, is it the total σ -algebra of the line in another guise? While the family of Borel sets is certainly very large, there are really wild subsets of the line out there that are not Borel. Constructing examples of such sets gets excessively technical and we won't pause to do so here; the Borel sets will be more than ample for our purposes and we will have no truck with the wilder sets. (For the morbidly curious, see the *Problems* in Chapter XI.)

The natural events of interest in continuous spaces are the intervals—and finite or denumerably infinite combinations of them—and to keep complexity in bounds it would be prudent to restrict attention to the smallest consistent family of sets containing the basic events of interest. In continuous spaces we will hence focus our attention resolutely on events in the σ -algebra of Borel sets.

EXAMPLE 2) Borel sets in \mathbb{R}^n . In the n -dimensional sample space \mathbb{R}^n we begin with the family \mathfrak{I} of finite unions of n -dimensional rectangles of the form $(a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_n, b_n]$. The Borel sets in Euclidean space \mathbb{R}^n are now the elements of the σ -algebra $\mathcal{B}(\mathbb{R}^n)$ generated by \mathfrak{I} . ▶

At the last the definition of the Borel σ -algebra was simple, deceptively so. While its provenance as the σ -algebra generated by intervals is clear, it is not at all clear what a generic Borel set looks like and whether familiar geometric objects are, in fact, Borel sets. The following section reassures the reader on this score.



9 A little point set topology

The reader will be familiar with the topology of the real line and, more expansively, finite-dimensional Euclidean space. In this setting she is no doubt familiar, at least at an intuitive level, with the concept of open sets. We begin with the real line.

DEFINITION 1 A subset \mathbb{A} of the real line is *open* if, for each $x \in \mathbb{A}$, there exists an open interval wholly contained in \mathbb{A} which has x as an interior point, that is, there exists $(a, b) \subseteq \mathbb{A}$ with $a < x < b$. A subset of the line is *closed* if its complement is open.

The union of any collection—finite, countably infinite, or even uncountably infinite—of open sets is open. This is easy to see as, if Λ is any index set and $\{\mathbb{A}_\lambda, \lambda \in \Lambda\}$ is any collection of open sets, then any point x in the union must lie in some set \mathbb{A}_λ and, as \mathbb{A}_λ is open, there exists an open interval wholly contained in \mathbb{A}_λ , and hence also in $\bigcup_{\lambda \in \Lambda} \mathbb{A}_\lambda$, that contains x as an interior point. This might suggest that open sets may have a very complicated structure. But a little thought should convince the reader that the elaboration of the structure of open sets on the line that follows is quite intuitive.

We need two elementary notions from the theory of numbers. A subset \mathbb{A} of real numbers is bounded if there are numbers a and b such that $a < x < b$ for every $x \in \mathbb{A}$. The largest of the lower bounds a is the greatest lower bound (or *infimum*) of \mathbb{A} and is denoted $\inf \mathbb{A}$; the smallest of the upper bounds b is the least upper bound (or *supremum*) of \mathbb{A} and is denoted $\sup \mathbb{A}$. The key fact that we will need is that every bounded set \mathbb{A} has a greatest lower bound and a least upper bound. (The reader who is not familiar with these notions will find them fleshed out in Section XI.1 in the Appendix.)

THEOREM 1 *Any open set on the line is the union of a countable number of pairwise disjoint open intervals and a fortiori every open set is a Borel set.*

PROOF: Suppose \mathbb{A} is an open subset of the real line, x a point in \mathbb{A} . Let \mathbb{I}_x be the union of all the open intervals wholly contained in \mathbb{A} for which x is an interior point. (It is clear that \mathbb{I}_x is non-empty as x has to be contained in at least one such interval.) Then $\mathbb{I}_x \subseteq \mathbb{A}$ for each $x \in \mathbb{A}$, whence $\bigcup_x \mathbb{I}_x \subseteq \mathbb{A}$; on the other hand, each $y \in \mathbb{A}$ lies in \mathbb{I}_y , hence also in $\bigcup_x \mathbb{I}_x$, and so $\bigcup_x \mathbb{I}_x \supseteq \mathbb{A}$. Thus, $\bigcup_x \mathbb{I}_x = \mathbb{A}$.

We now claim that \mathbb{I}_x is an open interval for each $x \in \mathbb{A}$ and, moreover, if $a = \inf \mathbb{I}_x$ and $b = \sup \mathbb{I}_x$ then $\mathbb{I}_x = (a, b)$. To see this, consider any point t with $a < t < x$. Then, by definition of infimum, there exists a point $s \in \mathbb{I}_x$ with $a < s < t$. But then s lies in some open interval wholly contained in \mathbb{A} and which contains x as an interior point. It follows that all points in the closed interval $[s, x]$ are contained in this interval and *a fortiori* so is t . This implies that all points $t \in (a, x)$ are contained in \mathbb{I}_x . An identical argument using the definition of supremum shows that all points $t \in (x, b)$ are also contained in \mathbb{I}_x . And, of course, it is patent that $x \in \mathbb{I}_x$. Thus, $(a, b) \subseteq \mathbb{I}_x$. But it is clear that $\mathbb{I}_x \subseteq (a, b)$ by the definition of the points a and b . It follows that, indeed, $\mathbb{I}_x = (a, b)$. Thus, we may identify \mathbb{I}_x as the *largest* open interval wholly contained in \mathbb{A} with x as an interior point.

Suppose now that x and y are distinct points in \mathbb{A} . Then either they engender the same open interval, $\mathbb{I}_x = \mathbb{I}_y$, or the intervals are not identical. If the latter is the case then the intervals \mathbb{I}_x and \mathbb{I}_y must in fact be disjoint. Else, if $\mathbb{I}_x \neq \mathbb{I}_y$ and $\mathbb{I}_x \cap \mathbb{I}_y \neq \emptyset$, then their union $\mathbb{I}_x \cup \mathbb{I}_y$ is an open interval larger than each of \mathbb{I}_x and \mathbb{I}_y , wholly contained in \mathbb{A} , and with both x and y as interior points. But this contradicts the maximality of both \mathbb{I}_x and \mathbb{I}_y . It follows that, for any two points x and y in \mathbb{A} , the corresponding open intervals \mathbb{I}_x and \mathbb{I}_y are either identical or disjoint.

Thus, as x sweeps through all the points of \mathbb{A} , the corresponding sets \mathbb{I}_x pick out a collection of pairwise disjoint open intervals. To each distinct interval in the collection we may assign a unique rational point inside it as an identifier, these rational values all distinct as the intervals making up the collection are disjoint. Thus, the intervals picked out as x sweeps through \mathbb{A} form a collection of disjoint intervals $\{\mathbb{I}_q, q \in \mathbb{Q}'\}$ indexed by some subset \mathbb{Q}' of the rational numbers. But the set of rationals may be placed in one-to-one correspondence with the natural numbers, that is to say, it is countable (the reader who does not know this fact will find a proof in Example XI.3.2), and so the disjoint family of open intervals $\{\mathbb{I}_q, q \in \mathbb{Q}'\}$ is countable.⁶ As $\mathbb{A} = \bigcup_q \mathbb{I}_q$, this means that \mathbb{A} can be written as a countable union of pairwise disjoint open intervals. As each open interval is a Borel set, manifestly so is \mathbb{A} . ▶

⁶Countability is intuitive but subtle. The reader new to such arguments will find other delicious examples in Section XI.3.

If the reader examines the proof we have just given critically, she will realise that it is an example of a “greedy” procedure which attempts to grab the largest open intervals possible at each step. A philosophically very different approach is to consider a succession of partitions of the line into smaller and smaller intervals and attempt to successively “fill up” the set \mathbb{A} . We need a little notation to put this idea into practice.

For each integer $m \geq 1$, let \mathcal{J}_m be the family of half-closed intervals of the form $(\frac{n}{2^m}, \frac{n+1}{2^m}]$ where n ranges over all integers. Then \mathcal{J}_m is a countable family of disjoint intervals whose union is \mathbb{R} . As the intervals in \mathcal{J}_{m+1} bisect those of \mathcal{J}_m , it follows by induction that, if $m < m'$, then $\mathcal{J}_{m'}$ is a *refinement* of \mathcal{J}_m in the sense that if $\mathbb{I} \in \mathcal{J}_m$ and $\mathbb{I}' \in \mathcal{J}_{m'}$, then either \mathbb{I}' is contained in \mathbb{I} or \mathbb{I}' and \mathbb{I} are disjoint.

For each m , we extract a subfamily \mathcal{J}_m of half-closed intervals from the family \mathcal{J}_m as follows. For $m = 1$, we include in \mathcal{J}_1 all the half-closed intervals \mathbb{I} in \mathcal{J}_1 whose closure is contained in \mathbb{A} . (The closure of a half-closed interval $(a, b]$ is the closed interval $[a, b]$.) For $m = 2$, we next include in \mathcal{J}_2 all the half-closed intervals \mathbb{I} in \mathcal{J}_2 whose closure is contained in \mathbb{A} and which are not contained in the half-closed intervals in \mathcal{J}_1 . For $m = 3$, we then include in \mathcal{J}_3 all the half-closed intervals \mathbb{I} in \mathcal{J}_3 whose closure is contained in \mathbb{A} and which are not contained in the half-closed intervals in either \mathcal{J}_1 or \mathcal{J}_2 . Proceeding in this fashion we recursively define the sequence $\{\mathcal{J}_m, m \geq 1\}$. Now let

$$\mathbb{A}' = \bigcup_{m \geq 1} \bigcup_{\mathbb{I} \in \mathcal{J}_m} \mathbb{I}, \quad (9.1)$$

that is to say, \mathbb{A}' is the countable union of the intervals in \mathcal{J}_m as m ranges over all values. As each interval \mathbb{I} on the right is contained in \mathbb{A} , it follows readily that $\mathbb{A}' \subseteq \mathbb{A}$. We now argue that in fact $\mathbb{A} \subseteq \mathbb{A}'$ as well.

Suppose x is any point in \mathbb{A} . Then there exists $\delta > 0$ such that the open interval $(x - \delta, x + \delta)$ is contained in \mathbb{A} . Let m be any integer m with $2^{-m} < \delta$. There then exists a unique integer n such that $n2^{-m} < x \leq (n+1)2^{-m}$. As $x - \delta < (n+1)2^{-m} - 2^{-m} = n2^{-m}$ and $x + \delta \geq n2^{-m} + 2^{-m} = (n+1)2^{-m}$, it follows that the interval $\mathbb{J} = (n2^{-m}, (n+1)2^{-m}]$ containing the point x is wholly contained in \mathbb{A} . If we select m to be the smallest integer with this property then $\mathbb{J} \in \mathcal{J}_m$ and it follows that x is contained in \mathbb{A}' . Thus, $\mathbb{A} \subseteq \mathbb{A}'$. We've hence shown that $\mathbb{A}' = \mathbb{A}$ and, as \mathbb{A}' is a countable union of intervals, it follows *a fortiori* that \mathbb{A} is indeed a Borel set.

Our second proof has taken a very different tack: it partitions space into finer and finer regions and essentially does a “water filling” of the set \mathbb{A} . While both proofs have their merits, the second one has the great virtue of being extendable to any number of dimensions, virtually without change.

Vector notation makes the demonstration practically transparent. Suppose $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ are points in the n -dimensional Euclidean space \mathbb{R}^n . By the vector inequality $\mathbf{x} < \mathbf{y}$ we mean that the componentwise inequalities $x_i < y_i$ hold for each i . We interpret the inequalities $\mathbf{x} \leq \mathbf{y}$, $\mathbf{x} > \mathbf{y}$, and $\mathbf{x} \geq \mathbf{y}$ likewise to mean that the corresponding componentwise inequalities are all simultaneously satisfied. Corresponding to open intervals in \mathbb{R} we may now define open rectangles in \mathbb{R}^n . Suppose \mathbf{a} and \mathbf{b} are points in \mathbb{R}^n with $\mathbf{a} < \mathbf{b}$. We define the *open rectangle* (\mathbf{a}, \mathbf{b}) to be the collection of points $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{a} < \mathbf{x} < \mathbf{b}$. The open rectangle (\mathbf{a}, \mathbf{b}) is hence simply the Cartesian product of open intervals $(\mathbf{a}, \mathbf{b}) = (a_1, b_1) \times \dots \times (a_n, b_n)$.

DEFINITION 2 A subset \mathbb{A} of \mathbb{R}^n is *open* if, for every point \mathbf{x} in \mathbb{A} , there exists an open

rectangle (\mathbf{a}, \mathbf{b}) wholly contained in \mathbb{A} and which contains the point x .⁷

Depending on whether we include the edges or not, we may now define rectangles of other types: the *closed rectangle* $[\mathbf{a}, \mathbf{b}]$ is the collection of points satisfying $\mathbf{a} \leq x \leq \mathbf{b}$, and the *half-closed rectangle* $(\mathbf{a}, \mathbf{b}]$ is the collection of points satisfying $\mathbf{a} < x \leq \mathbf{b}$. The *closure* of the open rectangle (\mathbf{a}, \mathbf{b}) (or of the half-closed rectangle $(\mathbf{a}, \mathbf{b}]$) is the closed rectangle $[\mathbf{a}, \mathbf{b}]$. We may allow one or more components of \mathbf{a} and \mathbf{b} to become infinite to obtain unbounded rectangles. We say that a rectangle (of any type) is *bounded* if all the components of \mathbf{a} and \mathbf{b} are finite.

THEOREM 2 *Any open set \mathbb{A} in \mathbb{R}^n may be expressed as the union of a countable number of bounded rectangles. It follows a fortiori that every open set in \mathbb{R}^n is a Borel set.*

PROOF: The water-filling construction extends to \mathbb{R}^n by the simple expedient of replacing half-closed intervals by half-closed rectangles. Let \mathcal{J}_m be the family of rectangles of side 2^{-m} obtained by taking Cartesian products of n half-closed intervals in the family \mathcal{J}_m . It is clear by the earlier argument that, if $m < m'$ then $\mathcal{J}_{m'}$ is a refinement of \mathcal{J}_m . Let \mathbb{A} be any open set in \mathbb{R}^n . We proceed as before to systematically accumulate half-closed rectangles inside \mathbb{A} . For $m = 1$, \mathcal{J}_1 is the collection of half-closed rectangles \mathbb{I} in \mathcal{J}_1 whose closure is contained in \mathbb{A} . For $m > 1$, \mathcal{J}_m is the collection of half-closed rectangles \mathbb{I} in \mathcal{J}_m whose closure is contained in \mathbb{A} and which are not contained within half-closed rectangles already accumulated in the families $\mathcal{J}_1, \dots, \mathcal{J}_{m-1}$. Figure 2 illustrates the process in two dimensions. Reusing notation, we define the set \mathbb{A}' as in (9.1), the sets \mathcal{J}_m now

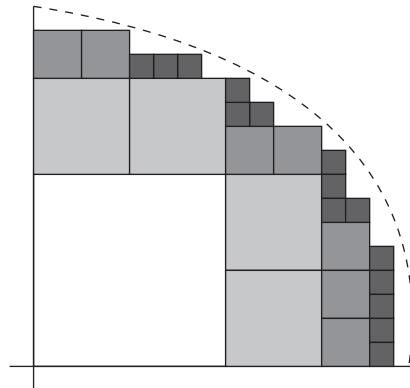


Figure 2: A “water-filling” of an open set by smaller and smaller open rectangles.

representing collections of half-closed rectangles. I will leave to the reader the minor tweaks in the argument needed to argue that $\mathbb{A}' = \mathbb{A}$. ▶

The basic link between analysis and topology in Euclidean spaces is through the open sets. But whereas the Borel σ -algebra is closed under countable unions, the

⁷Open sets are defined more traditionally in terms of open balls but our definition is equivalent to the standard definition and a little more convenient for our purposes.

family of open sets is topologically closed under *arbitrary* unions. That the family of Borel sets is sufficiently rich to include among its members such familiar topological objects as open sets, hence also closed sets, is reassuring. And, as we shall see in Section XII.6, this turns out to be of critical importance in the construction of a very rich class of probability experiments in the continuum.

10 Problems

Notation for generalised binomial coefficients will turn out to be useful going forward and I will introduce them here for ease of reference. As a matter of convention, for real t and integer k, we define

$$\binom{t}{k} = \begin{cases} \frac{t(t-1)(t-2)\cdots(t-k+1)}{k!} & \text{if } k \geq 0, \\ 0 & \text{if } k < 0. \end{cases}$$

Problems 1–5 deal with these generalised binomial coefficients.

1. *Pascal's triangle.* Prove that $\binom{t}{k-1} + \binom{t}{k} = \binom{t+1}{k}$.
2. If $t > 0$, show that $\binom{-t}{k} = (-1)^k \binom{t+k-1}{k}$ and hence that $\binom{-1}{k} = (-1)^k$ and $\binom{-2}{k} = (-1)^k (k+1)$ if $k \geq 0$.
3. Show that $\binom{1/2}{k} = (-1)^{k-1} \frac{1}{k} \binom{2k-2}{k-1} 2^{-2k+1}$ and $\binom{-1/2}{k} = (-1)^k \binom{2k}{k} 2^{-2k}$.
4. Prove *Newton's binomial theorem*

$$(1+x)^t = 1 + \binom{t}{1}x + \binom{t}{2}x^2 + \binom{t}{3}x^3 + \cdots = \sum_{k=0}^{\infty} \binom{t}{k}x^k,$$

the series converging for all real t whenever $|x| < 1$. Thence, if $t = n$ is any positive integer obtain the usual binomial formula $(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k$.

5. *Continuation.* With $t = -1$, the previous problem reduces to the *geometric series*

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \cdots$$

convergent for all $|x| < 1$. By integrating this expression termwise derive the *Taylor expansion of the natural logarithm*

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \cdots$$

convergent for $|x| < 1$. (Unless explicitly noted otherwise, our logarithms are always to the natural or "Napier" base e.) Hence derive the alternative forms

$$\begin{aligned} \log(1-x) &= -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \frac{1}{4}x^4 - \cdots, \\ \frac{1}{2} \log\left(\frac{1+x}{1-x}\right) &= x + \frac{1}{3}x^3 + \frac{1}{5}x^5 + \cdots. \end{aligned}$$

Problems 6–18 deal with sample spaces in intuitive settings.

6. A fair coin is tossed repeatedly. What is the probability that on the nth toss: (a) a head appears for the first time? (b) heads and tails are balanced? (c) exactly two heads have occurred?

7. Six cups and six saucers come in pairs, two pairs are red, two are white, and two are blue. If cups are randomly assigned to saucers find the probability that no cup is upon a saucer of the same colour.

8. *Birthday paradox.* In a group of n unrelated individuals, none born on a leap year, what is the probability that at least two share a birthday? Show that this probability exceeds one-half if $n \geq 23$. [Make natural assumptions for the probability space.]

9. *Lottery.* A lottery specifies a random subset R of r out of the first n natural numbers by picking one at a time. Determine the probabilities of the following events: (a) there are no consecutive numbers, (b) there is exactly one pair of consecutive numbers, (c) the numbers are drawn in increasing order. Suppose that you have picked your own random set of r numbers. What is the probability that (d) your selection matches R , (e) exactly k of your numbers match up with numbers in R ?

10. *Poker.* Hands at poker (see Example 3.4) are classified in the following categories: a *one pair* is a hand containing two cards of the same rank and three cards of disparate ranks in three other ranks; a *two pair* contains a pair of one rank, another pair of another rank, and a card of rank other than that of the two pairs; a *three-of-a-kind* contains three cards of one rank and two cards with ranks differing from each other and the rank of the cards in the triple; a *straight* contains five cards of consecutive ranks, not all of the same suit; a *flush* has all five cards of the same suit, not in sequence; a *full house* contains three cards of one rank and two cards of another rank; a *four-of-a-kind* contains four cards of one rank and one other card; and a *straight flush* contains five cards in sequence in the same suit. Determine their probabilities.

11. *Occupancy configurations, the Maxwell–Boltzmann distribution.* An occupancy configuration of n balls placed in r urns is an arrangement (k_1, \dots, k_r) where urn i has k_i balls in it. If the balls and urns are individually distinguishable, determine the number of distinguishable arrangements leading to a given occupancy configuration (k_1, \dots, k_r) where $k_i \geq 0$ for each i and $k_1 + \dots + k_r = n$. If balls are distributed at random into the urns determine thence the probability that a particular occupancy configuration (k_1, \dots, k_r) is discovered. These numbers are called the *Maxwell–Boltzmann statistics* in statistical physics. While it was natural to posit that physical particles were distributed in this fashion it came as a nasty jar to physicists to discover that no known particles actually behaved in accordance with common sense and followed this law.⁸

12. *Continuation, the Bose–Einstein distribution.* Suppose now that the balls are indistinguishable, the urns distinguishable. Let $A_{n,r}$ be the number of distinguishable arrangements of the n balls into the r urns. By lexicographic arrangement of the distinct occupancy configurations, show the validity of the recurrence $A_{n,r} = A_{n,r-1} + A_{n-1,r-1} + \dots + A_{1,r-1} + A_{0,r-1}$ with boundary conditions $A_{n,1} = 1$ for $n \geq 1$ and $A_{1,r} = r$ for $r \geq 1$. Solve the recurrence for $A_{n,r}$. [The standard mode of analysis is by a combinatorial trick by arrangement of the urns sequentially with sticks representing urn walls and identical stones representing balls. The alternative method suggested here provides a principled recurrence as a starting point instead.] The expression $1/A_{n,r}$ represents the probability that a given occupancy configuration is discovered assuming all occupancy configurations are equally likely. These are the *Bose–Einstein statistics* of statistical physics and have been found to apply to *bosons* such as photons, certain atomic

⁸S. Weinberg, *The Quantum Theory of Fields*. Cambridge: Cambridge University Press, 2000.

nuclei like those of the carbon-12 and helium-4 atoms, gluons which underlie the strong nuclear force, and W and Z bosons which mediate the weak nuclear force.

13. Continuation, the Fermi–Dirac distribution. With balls indistinguishable and urns distinguishable, suppose that the only legal occupancy configurations (k_1, \dots, k_r) are those where each k_i is either 0 or 1 only and $k_1 + \dots + k_r = n$, and suppose further that all legal occupancy configurations are equally likely. The conditions impose the constraint $n \leq r$ and prohibit occupancy configurations where an urn is occupied by more than one ball. Now determine the probability of observing a legal occupancy configuration (k_1, \dots, k_r) . These are the *Fermi–Dirac statistics* and have been found to apply to *fermions* such as electrons, neutrons, and protons.

14. Chromosome breakage and repair. Each of n sticks is broken into a long part and a short part, the parts jumbled up and recombined pairwise to form n new sticks. Find the probability (a) that the parts will be joined in the original order, and (b) that all long parts are paired with short parts.⁹

15. Spread of rumours. In a small town of n people a person passes a titbit of information to another person. A rumour is now launched with each recipient of the information passing it on to a randomly chosen individual. What is the probability that the rumour is told r times without (a) returning to the originator, (b) being repeated to anyone. *Generalisation:* redo the calculations if each person tells the rumour to m randomly selected people.

16. Keeping up with the Joneses. The social disease of keeping up with the Joneses may be parodied in this invented game of catch-up. Two protagonists are each provided with an n -sided die the faces of which show the numbers $1, 2, \dots, n$. Both social climbers start at the foot of the social ladder. Begin the game by having the first player roll her die and move as many steps up the ladder as show on the die face. The second player, envious of the progress of her rival, takes her turn next, rolls her die, and moves up the ladder as many steps as show on her die face. The game now progresses apace. At each turn, *whichever player is lower on the ladder* rolls her die and moves up as many steps as indicated on the die face. The game terminates at the first instant when both players end up on the same ladder step. (At which point, the two rivals realise the futility of it all and as the social honours are now even they resolve to remain friends and eschew social competition.) Call each throw of a die a turn and let N denote the number of turns before the game terminates. Determine the distribution of N . [Continued in Problem VII.7.]

17. The hot hand. A particular basketball player historically makes one basket for every two shots she takes. During a game in which she takes very many shots there is a period during which she seems to hit every shot; at some point, say, she makes five shots in a row. This is clearly evidence that she is on a purple patch (has a “hot hand”) where, temporarily at least, her chances of making a shot are much higher than usual, and so the team tries to funnel the ball to her to milk her run of successes as much as possible. Is this good thinking? Propose a model probability space for this problem and specify the event of interest.

⁹If sticks represent chromosomes broken by, say, X-ray irradiation, then a recombination of two long parts or two short parts causes cell death. See D. G. Catcheside, “The effect of X-ray dosage upon the frequency of induced structural changes in the chromosomes of *Drosophila Melanogaster*”, *Journal of Genetics*, vol. 36, pp. 307–320, 1938.



18. *Continuation, success run probability.* What is the probability that, in the normal course of things, the player makes five (or more) shots in a row somewhere among a string of 50 consecutive shot attempts? If the chance is small then the occurrence of a run of successes somewhere in an observed sequence of attempts would suggest either that an unlikely event has transpired or that the player was temporarily in an altered state (or, “in the zone”) while the run was in progress. Naturally enough, we would then attribute the observed run to a temporary change in odds (a hot hand) and not to the occurrence of an unlikely event. If, on the other hand, it turns out to be not at all unlikely that there will be at least one moderately long success run somewhere in a string of attempts, then one cannot give the hot hand theory any credence. [This problem has attracted critical interest¹⁰ and illustrates a surprising and counter-intuitive aspect of the theory of fluctuations. The analysis is elementary but not at all easy—Problems XV.20–27 provide a principled scaffolding on which problems of this type can be considered.]

The concluding Problems 19–34 are of a theoretical character.

19. *De Morgan’s laws.* Show that $(\bigcup_{\lambda} A_{\lambda})^c = \bigcap_{\lambda} A_{\lambda}^c$ and $(\bigcap_{\lambda} A_{\lambda})^c = \bigcup_{\lambda} A_{\lambda}^c$ where λ take values in an arbitrary index set Λ , possibly infinite.

20. Show that $(\bigcup_j A_j) \setminus (\bigcup_j B_j)$ and $(\bigcap_j A_j) \setminus (\bigcap_j B_j)$ are both subsets of $\bigcup_j (A_j \setminus B_j)$. When is there equality?

21. *σ -algebras containing two sets.* Suppose A and B are non-empty subsets of Ω and $A \cap B \neq \emptyset$. If $A \subseteq B$ determine the smallest σ -algebra containing both A and B . Repeat the exercise if $A \not\subseteq B$.

22. *Indicator functions.* Suppose A is any subset of a universal set Ω . The *indicator* for A , denoted 1_A , is the function $1_A(\omega)$ which takes value 1 when ω is in A and value 0 when ω is not in A . Indicators provide a very simple characterisation of the symmetric difference between sets. Recall that if A and B are subsets of some universal set Ω , we define $A \Delta B := (A \cap B^c) \cup (B \cap A^c) = (A \setminus B) \cup (B \setminus A)$. This is equivalent to the statement that $1_{A \Delta B} = 1_A + 1_B \pmod{2}$ where, on the right-hand side, the addition is modulo 2. Likewise, intersection has the simple representation, $1_{A \cap B} = 1_A \cdot 1_B \pmod{2}$. Verify the following properties of symmetric differences (and cast a thought to how tedious the verification would be otherwise): (a) $A \Delta B = B \Delta A$ (the commutative property), (b) $(A \Delta B) \Delta C = A \Delta (B \Delta C)$ (the associative property), (c) $(A \Delta B) \Delta (B \Delta C) = A \Delta C$, (d) $(A \Delta B) \Delta (C \Delta D) = (A \Delta C) \Delta (B \Delta D)$, (e) $A \Delta B = C$ if, and only if, $A = B \Delta C$, (f) $A \Delta B = C \Delta D$ if, and only if, $A \Delta C = B \Delta D$. In view of their indicator characterisations, it now becomes natural to identify symmetric difference with “addition”, $A \oplus B := A \Delta B$, and intersection with “multiplication”, $A \otimes B := A \cap B$.

23. *Why is the family of events called an algebra?* Suppose \mathcal{F} is a non-empty family of subsets of a universal set Ω that is closed under complementation and finite unions.

¹⁰T. Gilovich, R. Vallone, and A. Tversky, “The hot hand in basketball: On the misperception of random sequences”, *Cognitive Psychology*, vol. 17, pp. 295–314, 1985. Their conclusion? That the hot hand theory is a widespread cognitive illusion affecting all beholders, players, coaches, and fans. Public reaction to the story was one of disbelief. When the celebrated cigar-puffing coach of the Boston Celtics, Red Auerbach, was told of Gilovich and his study, he grunted, “Who is this guy? So he makes a study. I couldn’t care less”. Auerbach’s quote is reported in D. Kahneman, *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux, 2011, p. 117.

First argue that \mathcal{F} is closed under intersections, set differences, and symmetric differences. Now, for A and B in \mathcal{F} , define $A \otimes B := A \cap B$ and $A \oplus B := A \Delta B$. Show that equipped with the operations of “addition” (\oplus) and “multiplication” (\otimes), the system \mathcal{F} may be identified as an algebraic system in the usual sense of the word. That is, addition is commutative, associative, and has \emptyset as the zero element, i.e., $A \oplus B = B \oplus A$, $A \oplus (B \oplus C) = (A \oplus B) \oplus C$, and $A \oplus \emptyset = A$, and multiplication is commutative, associative, distributes over addition, and has Ω as the unit element, i.e., $A \otimes B = B \otimes A$, $A \otimes (B \otimes C) = (A \otimes B) \otimes C$, $A \otimes (B \oplus C) = (A \otimes B) \oplus (A \otimes C)$, and $A \otimes \Omega = A$.

24. Show that a σ -algebra \mathcal{F} is closed under countable intersections.

25. If $\mathcal{A} = \{A_1, \dots, A_n\}$ is a finite family of sets what is the maximum number of elements that $\sigma(\mathcal{A})$ can have?

26. For any $a < b$ show that the open interval (a, b) may be obtained via a countable number of operations on half-closed intervals by showing that (a, b) may be represented as the countable union of the half-closed intervals $(a, b - (b - a)/n]$ as n varies over all integers ≥ 1 . Likewise show how the closed interval $[a, b]$ and the reversed half-closed interval $[a, b)$ may be obtained from half-closed intervals by a countable number of operations. Show how to generate the singleton point $\{a\}$ by a countable number of operations on half-closed intervals.

27. Continuation. Show that intervals of any type are obtainable by countable operations on intervals of a given type.

28. Increasing sequences of sets. Suppose $\{A_n, n \geq 1\}$ is an increasing sequence of events, $A_n \subseteq A_{n+1}$ for each n . Let $A = \bigcup_{n \geq 1} A_n$. Show that $P(A_n) \rightarrow P(A)$ as $n \rightarrow \infty$.

29. Decreasing sequences of sets. Suppose $\{B_n, n \geq 1\}$ is a decreasing sequence of events, $B_n \supseteq B_{n+1}$ for each n . Let $B = \bigcap_{n \geq 1} B_n$. Show that $P(B_n) \rightarrow P(B)$ as $n \rightarrow \infty$.

30. Subadditivity. Suppose that P is a set function on the σ -algebra \mathcal{F} satisfying Axioms 1, 2, and 3 of probability measure. If it is true that for all sequences $B_1, B_2, \dots, B_n, \dots$ of sets in \mathcal{F} satisfying $A \subseteq \bigcup_n B_n$ we have $P(A) \leq \sum_n P(B_n)$ then show that P satisfies Axiom 4. It is frequently easier to verify the continuity axiom by this property.

 **31. A semiring of sets.** Suppose Ω is the set of all rational points in the unit interval $[0, 1]$ and let \mathcal{A} be the set of all intersections of the set Ω with arbitrary open, closed, and half-closed subintervals of $[0, 1]$. Show that \mathcal{A} has the following properties: (a) $\emptyset \in \mathcal{A}$; (b) \mathcal{A} is closed under intersections; and (c) if A_1 and A are elements of \mathcal{A} with $A_1 \subseteq A$ then A may be represented as a finite union $A = A_1 \cup A_2 \cup \dots \cup A_n$ of pairwise disjoint sets in \mathcal{A} , the given set A_1 being the first term in the expansion. A family of sets with the properties (a), (b), and (c) is called a *semiring*.

 **32. Continuation.** For each selection of $0 \leq a \leq b \leq 1$, let $A_{a,b}$ be the set of points obtained by intersecting Ω with any of the intervals (a, b) , $[a, b]$, $[a, b)$, or $(a, b]$. Define a set function Q on the sets of \mathcal{A} by the formula $Q(A_{a,b}) = b - a$. Show that Q is additive but not countably additive, hence not continuous. [Hint: Although $Q(\Omega) = 1$, Ω is a countable union of single-element sets, each of which has Q -measure zero.]

 **33. Continuation.** Suppose $A \in \mathcal{A}$ and $A_1, A_2, \dots, A_n, \dots$ is a sequence of pairwise disjoint subsets of A , all belonging to \mathcal{A} . Show that $\sum_n Q(A_n) \leq Q(A)$.

 **34. Continuation.** Show that there exists a sequence $B_1, B_2, \dots, B_n, \dots$ of sets in \mathcal{A} satisfying $A \subseteq \bigcup_n B_n$ but $Q(A) > \sum_n Q(B_n)$. This should be contrasted with Problem 30.

II

Conditional Probability

The simplest modification to a probability measure arises when information about the outcome of a chance experiment is provided obliquely by specifying something about the nature of the outcome without explicitly identifying it. An attempt to codify the impact of such side information leads to the construct of conditional probability. While elementary, conditioning arguments are pervasive and their appearance in problems can be subtle as the reader will find in the applications scattered through this chapter.

C 1, 3, 7
A 2, 4–6, 8–10

1 Chance domains with side information

The character of the definition of conditional probability is best illustrated by settings in our common experience.

EXAMPLES: 1) *Return to coin tossing.* Suppose a (fair) coin is tossed thrice. Identifying the results of the tosses sequentially, the sample space is identified as the set of eight elements

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

each of which has equal probability 1/8 of occurrence. Let A be the event that the first toss is a head. It is clear that the probability of A is 4/8 = 1/2.

Suppose now that one is informed that the outcome of the experiment was exactly one head. What is the probability of A *conditioned upon* this information? Let H be the event that exactly one head occurs. Then H consists of the sample points HTT , THT , and TTH . If exactly one head occurs then the outcome must be one of the three elements of H each of which performance is equally likely to have been the observed outcome. The event A can then occur if, and only if, the observed outcome was HTT . Consequently, the probability of A given that H has occurred is now 1/3. Side information about the outcome of the experiment in the form of the occurrence of H affects the projections of whether the outcomes comprising A could have occurred.

2) *Dice.* Suppose two six-sided dice are thrown. The probability of the event A that at least one six is recorded is then easily computed to be $1 - 25/36 = 11/36$. If one is informed, however, that the sum of the face values is 8 then the possible outcomes of the experiment reduce from 36 pairs of integers (i, j) with $1 \leq i, j \leq 6$ to the outcomes $\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$ only, each of which is equally likely to occur. The probability that at least one six is recorded, conditioned on the sum of face values being 8, is then $2/5$.

3) *Families.* A family has two children. If one of them is known to be a boy, what is the probability that the other is also a boy? Unprepared intuition may suggest a probability of one-half. But consider: listing the genders of the two children, elder first, the sample space for this problem may be considered to be $\Omega = \{bb, bg, gb, gg\}$ with the natural assignment of probability $1/4$ to each of the four possibilities. If it is known that one child is a boy, the sample space reduces to the three equally likely possibilities $\{bb, bg, gb\}$. Only one of the outcomes, bb , in the reduced space is identified with the event that the other child is a boy and so, given that one child is a boy, the probability that the other is also a boy is $1/3$. Again, side information about the outcome of the experiment in the form of the occurrence of an auxiliary event affects event probabilities. ►

In these examples, the sample space is finite with all sample points possessed of the same probability (the combinatorial case) and side information about the outcome of the experiment is codified in the form of the occurrence of an auxiliary event H. Based on this information we may consider the original sample space Ω to be reduced in size to a new sample space comprised of the sample points in H alone as knowledge of the occurrence of H precludes any point in H^c . It is natural then to consider that, conditioned upon the occurrence of H, the probability that an event A has occurred is modified to be the proportional number of points shared by A and H out of the totality of points in H and we express this in notation in the form

$$P(A | H) = \frac{\text{card}(A \cap H)}{\text{card } H} = \frac{\# \text{ of elements in } A \cap H}{\# \text{ of elements in } H}.$$

More generally, in an abstract space Ω that is not necessarily finite, the proper expression of the idea of the conditional probability of the occurrence of A given that H has occurred relies upon the natural normalisation of the points in $A \cap H$ vis à vis the totality of points in H which now comprises the effective sample space for the problem.

DEFINITION Let A be any event in a probability space and H any event of non-zero probability. The *conditional probability of A given that H has occurred* (or, in short, the *probability of A given H*) is denoted $P(A | H)$ and defined by

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

The conditional probability is undefined if the event H has zero probability.

It should be borne in mind that native event probabilities may be modified dramatically by side information. A simple numerical example may serve to reinforce this point.

EXAMPLE 4) Consider the sample space $\Omega = \{1, \dots, 10\}$ corresponding to a ten-sided die where every singleton event has equal probability, $p_k = P\{k\} = 1/10$ for $1 \leq k \leq 10$. If $A = \{2, 3, 4, 5, 6\}$ and $H = \{4, 5, 6, 7, 8, 9\}$, then $P(A) = 5/10$, $P(H) = 6/10$, and $P(A \cap H) = 3/10$, whence it follows that

$$P(A | H) = \frac{3}{10} / \frac{6}{10} = \frac{1}{2} = P(A) \text{ and } P(H | A) = \frac{3}{10} / \frac{5}{10} = \frac{3}{5} = P(H).$$

The events A and H appear to be “independent”: the frequency of occurrence of A does not appear to be affected by conditioning on the occurrence of H , and vice versa. If, on the other hand, we consider the event $B = \{2, 4, 6, 8, 10\}$ of probability $P(B) = 5/10$, it is easy to see that $P(A \cap B) = 3/10$ and $P(H \cap B) = 3/10$. It follows now that

$$P(A | B) = \frac{3}{10} / \frac{5}{10} = \frac{3}{5} > P(A) \text{ and } P(H | B) = \frac{3}{10} / \frac{5}{10} = \frac{3}{5} = P(H).$$

Finally, consider the event $C = \{1, 7, 8, 9\}$ of probability $P(C) = 4/10$. Then $P(A \cap C) = 0$ and $P(H \cap C) = 3/10$, so that

$$P(A | C) = 0 < P(A) \text{ and } P(H | C) = \frac{3}{10} / \frac{4}{10} = \frac{3}{4} > P(H).$$

In other words, *conditioning provides information that can affect event probabilities in unexpected ways*. ▶

Side information in the form of an event H effectively modifies the probability measure $P(\cdot)$, replacing it by a new probability measure,

$$P_H: A \mapsto P(A | H),$$

on the underlying σ -algebra. The new measure $P_H(\cdot)$ places all its mass in H , $P_H(A) = 0$ if $A \in H^c$ and $P_H(H) = 1$. It is easy to verify by stepping through the axioms that $P_H(\cdot)$ is indeed a probability measure on the space Ω .

- ① The new measure is positive, $P_H(A) \geq 0$, as is trite.
- ② The measure is properly normalised, $P_H(\Omega) = 1$, as is also trivial to verify.
- ③ If A and B are disjoint events then

$$\begin{aligned} P(A \cup B | H) &= \frac{P((A \cup B) \cap H)}{P(H)} = \frac{P((A \cap H) \cup (B \cap H))}{P(H)} \\ &= \frac{P(A \cap H)}{P(H)} + \frac{P(B \cap H)}{P(H)} = P(A | H) + P(B | H). \end{aligned}$$

It follows that $P_H(A \cup B) = P_H(A) + P_H(B)$ whenever $A \cap B = \emptyset$ and the new measure $P_H(\cdot)$ is additive.

Conditional Probability

- ④ Suppose $\{A_n, n \geq 1\}$ is a decreasing sequence of events, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$, with $\bigcap_n A_n = \emptyset$. Then $0 \leq P(A_n | H) = P(A_n \cap H)/P(H) \leq P(A_n)/P(H) \rightarrow 0$ by monotonicity and continuity of the original probability measure $P(\cdot)$. It follows that $P_H(A_n) \rightarrow 0$ and the new measure $P_H(\cdot)$ is continuous.

Repeated application of the definition of conditioning to three events, say A , B , and C , leads to the identity

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

Of course, it is tacitly assumed that $P(B \cap C) \neq 0$.

An iterative application of this process leads to the chain rule for events which is useful enough to be worth enshrining as a theorem.

THE CHAIN RULE FOR CONDITIONAL PROBABILITIES *Let A_1, \dots, A_n be events whose intersection has strictly positive probability. Then*

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_n) &= P(A_1 | A_2 \cap \dots \cap A_n) \times P(A_2 | A_3 \cap \dots \cap A_n) \\ &\quad \times \dots \times P(A_{n-1} | A_n) \times P(A_n). \end{aligned}$$

EXAMPLES: 5) In Example 4 what can we say about the conditional probability of H given A and B ? Via the chain rule,

$$P(H | A \cap B) = \frac{P(A \cap B \cap H)}{P(A | B) P(B)} = \frac{2}{10} / \left(\frac{3}{5} \cdot \frac{1}{2} \right) = \frac{2}{3},$$

and we observe that

$$P(H) = P(H | A) = P(H | B) < P(H | A \cap B).$$

The moral, again, is that conditioning affects probabilities in unexpected ways.

6) *Bridge.* A bridge hand is the distribution of a standard pack of 52 playing cards among four players, traditionally called North, South, East, and West, 13 cards to a player. Aces in the various suits play an important rôle in bridge. Accordingly, it is of interest to determine the probability of a balanced hand in aces, that is to say, the probability that each player possesses exactly one ace.

Let E_{\spadesuit} be the event that some player gets the ace of spades, $E_{\spadesuit, \heartsuit}$ the event that the ace of spades and the ace of hearts go to two distinct players, $E_{\spadesuit, \heartsuit, \diamondsuit}$ the event that the aces of spades, hearts, and diamonds go one apiece to three distinct players, and $E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit}$ the event that each player gets one ace. It is clear that each of these events is implied by the succeeding event,

$$E_{\spadesuit} \supseteq E_{\spadesuit, \heartsuit} \supseteq E_{\spadesuit, \heartsuit, \diamondsuit} \supseteq E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit},$$

these inclusion relations implying tritely that

$$E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit} = E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit} \cap E_{\spadesuit, \heartsuit, \diamondsuit} \cap E_{\spadesuit, \heartsuit} \cap E_{\spadesuit}.$$

We may hence leverage the chain rule to obtain the expression

$$P(E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit}) = P(E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit} | E_{\spadesuit, \heartsuit, \diamondsuit}) P(E_{\spadesuit, \heartsuit, \diamondsuit} | E_{\spadesuit, \heartsuit}) P(E_{\spadesuit, \heartsuit} | E_{\spadesuit}) P(E_{\spadesuit}).$$

Now it is clear that the ace of spades must end up with one player or the other, whence

$$P(E_{\spadesuit}) = 1.$$

Conditioned upon the ace of spades being with one of the four players, the ace of hearts can be in the possession of a different player if, and only if, it is among the 39 cards held by the players who do not have the ace of spades. As there are 51 remaining locations where the ace of hearts could be placed, we have

$$P(E_{\heartsuit} | E_{\spadesuit}) = 39/51.$$

Now conditioned upon the aces of spades and hearts distributed between two different players, the ace of diamonds can be in the possession of yet a third player if, and only if, it is among the 26 cards held by the remaining two players. And thus, as there are 50 remaining locations where the ace of diamonds could be placed, we have

$$P(E_{\diamondsuit} | E_{\spadesuit, \heartsuit}) = 26/50.$$

And, arguing in this vein, conditioned upon the aces of spades, hearts, and diamonds being in the possession of three distinct players, the ace of clubs is in the possession of the last player with probability

$$P(E_{\clubsuit} | E_{\spadesuit, \heartsuit, \diamondsuit}) = 13/49$$

as it must lie in his hand of 13 cards and there are 49 residual positions it could have been placed in among the four hands. Pooling the results, we obtain

$$P(E_{\spadesuit, \heartsuit, \diamondsuit, \clubsuit}) = \frac{13 \cdot 26 \cdot 39}{49 \cdot 50 \cdot 51} \approx 0.105.$$

Approximately 10.5% of bridge hands have balanced aces. ▶

In our definitions hitherto, we have viewed conditional probabilities merely as the artefacts of available side information about the probabilistic process: the underlying probabilities were fixed and immutable and knowledge about the outcome of the experiment was codified in the form of the occurrence of an event. This information could then be used to compute conditional probabilities. In typical applications, however, the process is reversed: in many settings it turns out to be more natural to specify conditional probabilities from which one can infer the underlying probabilities. We will see this illustrated repeatedly in examples.

2 Gender bias? Simpson's paradox

Is there pervasive gender bias in graduate admissions to elite colleges? In the fall of 1973, the Graduate Division of the University of California, Berkeley, admitted 44% of male applicants and 35% of female applicants. The apparent gender bias prompted a study of graduate admissions at the university.¹ The following hypothetical data for two unusual departments, Social Warfare and Machismatics, in a make-believe university are taken from the Berkeley study and illustrate the perils in making snap judgements.

	Social Warfare		Machismatics	
	Women	Men	Women	Men
Admitted	20	1	19	100
Rejected	180	19	1	100

Table 1: Admissions data for two unusual departments.

Comparing admission ratios by department, it is clear that in the Department of Social Warfare the admission rate of 1/10 for women is twice that of the admission rate 1/20 for men, while in the Department of Machismatics the admission rate of 19/20 for women far outstrips the admission rate of 1/2 for men. While the significance of the results may be debated, it is clear that the ratios at any rate suggest an admissions edge for women. The picture changes dramatically, however, if the situation is considered in entirety for both departments put together. Now the cumulative admissions rate for women is 39/220 or about one in five while the men enjoy an admissions rate of 101/220 or about one in two. Viewed holistically, the men seem to have a decided edge in admissions though the individual parts show a fairly pronounced edge for women. How to read this riddle?

Suppose a member of the aspiring graduates is selected at random from the pool of 440 applicants. Let A be the event that the student is admitted, B the event that the student is a woman, and C the event that the student is an applicant to the Department of Social Warfare. In notation then, the apparently paradoxical situation before us is expressed via the observation that while

$$\mathbf{P}(A | B \cap C) \geq \mathbf{P}(A | B^c \cap C) \text{ and } \mathbf{P}(A | B \cap C^c) \geq \mathbf{P}(A | B^c \cap C^c), \quad (2.1)$$

yet, when the situations are combined,

$$\mathbf{P}(A | B) \leq \mathbf{P}(A | B^c), \quad (2.2)$$

¹P. J. Bickel, E. A. Hammel, J. W. O'Connell, "Sex-bias in graduate education: Data from Berkeley", *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

and the inequality appears to unfairly change direction.

<i>Basic events</i>	$A \cap B \cap C$	$A^c \cap B \cap C$	$A \cap B \cap C^c$	$A^c \cap B \cap C^c$
<i>Probabilities</i>	a	b	c	d
<i>Basic events</i>	$A \cap B^c \cap C$	$A^c \cap B^c \cap C$	$A \cap B^c \cap C^c$	$A^c \cap B^c \cap C^c$
<i>Probabilities</i>	e	f	g	h

Table 2: Probabilities of basic events.

To understand the conditions under which this perplexing state of affairs can persist, consider the space determined by three events A , B , and C and introduce notation for the probabilities of the eight elemental events induced by A , B , and C as shown in Table 2. The Venn diagram shown in Figure 1 may help clarify the division of spoils.

Simple applications of additivity now show that for (2.1,2.2) to hold we must satisfy all three inequalities

$$\frac{a}{a+b} \geq \frac{e}{e+f}, \quad \frac{c}{c+d} \geq \frac{g}{g+h}, \quad \frac{a+c}{a+c+b+d} \leq \frac{e+g}{e+g+f+h}.$$

Simplifying, it becomes clear that the inequalities (2.1,2.2) can hold jointly if, and only if, $af \geq be$, $ch \geq dg$, and $(a+c)(f+h) \leq (b+d)(e+g)$, subject to the positivity and normalisation conditions, $a, b, \dots, h \geq 0$ and $a + b + \dots + h = 1$. Geometrically speaking, these conditions are equivalent to the requirement that there exist rectangles R_1 and R_2 as shown in Figure 2 with $\text{area}(A_1) \geq \text{area}(A_2)$, $\text{area}(B_1) \geq \text{area}(B_2)$, and $\text{area}(R_1) \leq \text{area}(R_2)$.

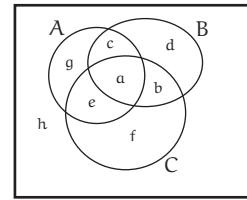


Figure 1: The eight basic events engendered by A , B , C .

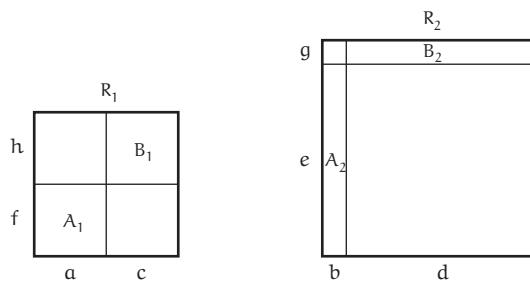


Figure 2: A geometrical illustration of Simpson's paradox.

Many solutions can be written down on a more or less *ad hoc* basis. For instance, we may specify $a = c = h = f = 3/30$, $b = g = 1/30$, and

$d = e = 8/30$, the situation shown in the figure. Intuition can be salvaged once the algebraic basis for the apparent incongruity is understood: the paradoxical nature of the result is explained by the fact that in the given example there is a tendency for women to apply in larger numbers to a department where both sexes are admitted in small numbers. While the result has been known in various sociological and statistical settings since early in the twentieth century, interest in it was sparked in 1951 by E. H. Simpson, for whom the apparent paradox is named, with a variety of amusing and surprising examples where this situation occurs.²

3 The theorem of total probability

The countable additivity of probability measure manifests itself under conditioning in an extremely useful, if elementary, form. Some terminology first. We say that a sequence of events, $\{A_j, j \geq 1\}$, forms a *measurable partition* of Ω if $A_j \cap A_k = \emptyset$ if $j \neq k$ and $\bigcup_j A_j = \Omega$, that is to say, the event sequence is pairwise disjoint and covers the sample space.

THEOREM OF TOTAL PROBABILITY *Suppose $\{A_j, j \geq 1\}$ is a measurable partition of Ω and $P(A_j) > 0$ for each j . Then, for every event H , we have*

$$P(H) = \sum_{j \geq 1} P(A_j \cap H) = \sum_{j \geq 1} P(H | A_j) P(A_j).$$

PROOF: The result follows from the trite observation $H = \bigcup_j (A_j \cap H)$ (as $H \subseteq \Omega = \bigcup_j A_j$). Countable additivity of probability measure completes the proof as $\{A_j \cap H, j \geq 1\}$ is a sequence of pairwise disjoint events. ►

The simplest non-trivial measurable partition of the sample space arises by splitting up Ω into an event and its complement. Thus, if A is any event in the probability space then $\{A, A^c\}$ is a measurable partition of Ω . In consequence,

$$P(H) = P(H | A) P(A) + P(H | A^c) P(A^c)$$

for any pair of events H and A where, to obviate trivialities, we suppose that both A and A^c have strictly positive probability [equivalently, $0 < P(A) < 1$]. This apparently banal observation turns out to be quite amazingly useful and will in itself pay the rent as we see in the examples that follow.

EXAMPLES: 1) *Let's make a deal!* A popular game show called *Let's Make a Deal* made its debut on the NBC Television Network on December 30, 1963. In the

²E. H. Simpson, "The interpretation of interaction in contingency tables", *Journal of the Royal Statistical Society, Series B*, vol. 13, pp. 238–241, 1951.

simplest version of the game, a prize is placed behind one of three closed doors and gag items of little or no value (called *zonks* in the show) are placed behind the other two doors, also closed. Contestants in the show are aware of the fact that one of the three doors conceals a prize of some value but do not know behind which door it is. To begin, a contestant is asked to select a door (but not open it). Once the selection is made, the moderator, the charismatic Monty Hall, opens one of the remaining two doors and displays a zonk behind it. (As at least one of the remaining two doors must conceal a zonk, it is always possible for him to select a door concealing a zonk from the remaining two.) The contestant is now given the option of either staying with her original door selection or switching allegiance to the remaining door in the hope of getting the prize. Once she has made her decision, she opens her final door selection and can take possession of whatever is behind the door. What should her strategy be to maximise her chances of winning the prize?

One is inclined to reflexively believe that the probability of securing a prize is 1/2 whether the contestant stays with her original choice or picks the remaining door. After all, the prize is behind one of the two unopened doors and it seems eminently reasonable to model the probability of it being behind either door as one-half. A more careful examination shows, however, that this facile argument is flawed and that the contestant stands to gain by always switching to the remaining door.

What is the sample space for the problem? We may begin by supposing that all six permutations of the prize and the two zonks are equally likely. Writing \star for the prize and α and β for the two zonks, the possible arrangements, listed in order by door, are shown in Figure 3. Write A for the event that the

Prize behind first door	Prize behind second door	Prize behind third door
$\star \alpha \beta$	$\star \beta \alpha$	$\alpha \star \beta$

Figure 3: Arrangements of prize and zonks in Let's Make a Deal.

prize is behind the door initially selected by the contestant and write B for the event that the prize is behind the remaining door after Monty Hall has made his selection. By the symmetry inherent in the problem it makes no matter which door the contestant picks initially and for definiteness let us suppose that she picks the first door. Then we may identify $A = \{\star \alpha \beta, \star \beta \alpha\}$. If A occurs then, regardless of the strategy used by Monty Hall for selecting a door to open, the remaining door will conceal a zonk (as the second and third doors both conceal zonks). If, however, A does not occur, then the prize is behind the second or third door and Monty Hall is compelled to select and open the unique door of the two that does not conceal the prize. But then the residual door must conceal

the prize. We may hence identify $B = A^c = \{\alpha * \beta, \beta * \alpha, \alpha \beta *, \beta \alpha *\}$, all these outcomes of the experiment resulting in the prize lying behind the remaining door after Monty Hall has made his selection. It follows that $P(A) = 1/3$ and $P(B) = 2/3$. The analysis is unaffected by the particular initial choice of the contestant and, in consequence, the strategy of switching, if followed religiously, will unearth the prize two-thirds of the time.

A number of variations on the theme can be proposed based on the moderator's strategy. This example has attained a measure of notoriety due to its apparently paradoxical nature but conditioning on the initial choice makes clear where the fallacy in logic lies.

2) *The ballot problem; or, leading from tape to tape.* In an election between two candidates \mathfrak{A} and \mathfrak{B} , suppose $m + n$ ballots are cast in all, \mathfrak{A} getting n votes and \mathfrak{B} getting m votes. Suppose $m < n$ so that \mathfrak{A} wins the election. If the $n + m$ ballots are cast sequentially in random order, what is the probability that \mathfrak{A} leads \mathfrak{B} throughout the count? Problems of this stripe which may be modelled as a succession of trials are particularly amenable to recursive arguments by conditioning on the first or the last step of the sequence. I will illustrate the method here by conditioning on the result of the final ballot.

We begin by supposing, naturally enough, that all $(n + m)!$ arrangements of ballot order are equally likely. We are interested in the probability of the event A that \mathfrak{A} leads throughout the count. Now, for any ordering of the ballots, there are one of two eventualities: the final ballot is cast either for \mathfrak{A} or for \mathfrak{B} . Writing F for the event that the final ballot is cast for \mathfrak{A} , by total probability, we obtain $P(A) = P(A | F)P(F) + P(A | F^c)P(F^c)$, and, if we have discovered the key to the problem, the probabilities on the right should devolve into simple expressions. We begin with a consideration of the final ballot cast. The number of arrangements for which the last ballot is one of the n cast in favour of \mathfrak{A} is $n(n + m - 1)!$ and so the probability that the final ballot is cast for \mathfrak{A} is given by

$$P(F) = \frac{n(n + m - 1)!}{(n + m)!} = \frac{n}{n + m}, \text{ whence also } P(F^c) = \frac{m}{n + m}.$$

Now, the probability that \mathfrak{A} leads at every step of the count is determined completely by n and m and, accordingly, to keep the dependence firmly in view, write $P_{n,m} = P(A)$. Given that the last ballot is cast in favour of \mathfrak{A} then, of the first $n + m - 1$ ballots cast, \mathfrak{A} receives $n - 1$, \mathfrak{B} receives m , and, if the event A is to occur, then \mathfrak{A} must have led through every one of the first $n + m - 1$ steps (and hence also through the last step). As all arrangements of the first $n + m - 1$ ballots are equally likely, the conditional probability that \mathfrak{A} leads through the first $n + m - 1$ steps given that he gets the last ballot is given hence, in our notation, by $P(A | F) = P_{n-1,m}$. Likewise, if \mathfrak{B} receives the final ballot then, of the first $n + m - 1$ ballots cast, \mathfrak{A} receives n , \mathfrak{B} receives $m - 1$, and the conditional

probability that \mathfrak{A} leads every step of the way is given by $P(A | F^c) = P_{n,m-1}$. We have hence obtained the elegant recurrence

$$P_{n,m} = P_{n-1,m} \cdot \frac{n}{n+m} + P_{n,m-1} \cdot \frac{m}{n+m} \quad (1 \leq m < n). \quad (3.1)$$

For the boundary conditions it is clear that $P_{n,n} = 0$ for $n \geq 1$ and $P_{n,0} = 1$ for $n \geq 0$. Starting from the boundaries it is now a simple matter to recursively generate successive values for these probabilities as shown in Table 3.

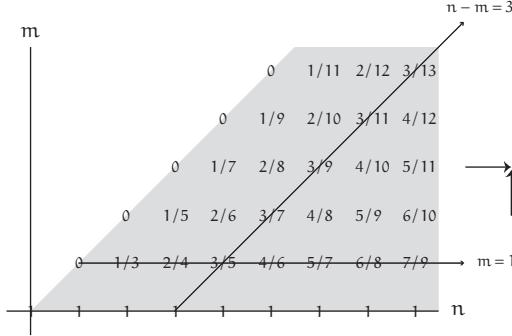


Table 3: Probabilities for the ballot problem.

Analytical solutions are preferable, when available, and one is within reach here. The row of values when $m = 1$ is suggestive and it is not hard to verify the natural guess that $P_{n,1} = (n-1)/(n+1)$ analytically. As $P_{n,0} = 1$ for $n \geq 1$, by churning the recurrence crank we obtain

$$\begin{aligned} P_{n,1} &= P_{n-1,1} \cdot \frac{n}{n+1} + \frac{1}{n+1} = \left[P_{n-2,1} \cdot \frac{n-1}{n} + \frac{1}{n} \right] \frac{n}{n+1} + \frac{1}{n+1} \\ &= P_{n-2,1} \cdot \frac{n-1}{n} \cdot \frac{n}{n+1} + \frac{2}{n+1} = \dots = P_{1,1} \cdot \frac{2}{3} \cdot \frac{3}{4} \cdots \frac{n}{n+1} + \frac{n-1}{n+1} = \frac{n-1}{n+1} \end{aligned}$$

as anticipated. Is there a direct combinatorial pathway to this answer?

The general solution is not much harder. If the reader squints along the sight lines of the 45° lines $n - m = \text{constant}$ in the table of numerical values for the probabilities $P_{n,m}$, she will begin to entertain the suspicion that

$$P_{n,m} = \frac{n-m}{n+m} \quad (1 \leq m \leq n)$$

and induction readily bears this out. As induction hypothesis, suppose that $P_{n-1,m} = (n-m-1)/(n+m-1)$ and $P_{n,m-1} = (n-m+1)/(n+m-1)$. It is now a simple algebraic matter to verify that the right-hand side of (3.1) reduces to

$$\frac{n-m-1}{n+m-1} \cdot \frac{n}{n+m} + \frac{n-m+1}{n+m-1} \cdot \frac{m}{n+m} = \frac{n-m}{n+m}$$

to complete the induction.

The result carries some surprise value. Of two-party elections where the winning party has a majority of 55% to 45%, the winning candidate will lead from start to finish in fully 10% of all such cases, a result the reader may find surprising. The problem was first solved by W. A. Whitworth in 1878³ using a recursive argument of the kind considered here; it has since been vastly generalised. This example turns out to have a particular importance in the theory of random walks and we return to it from a fresh vantage point in the Ballot Theorem of Section VIII.4.

3) *Guessing strategies; or, there's no free lunch.* The numbers $1, \dots, n$ are randomly shuffled to obtain the permuted sequence π_1, \dots, π_n . Any *a priori* guess at the location of the number 1 will then have a $1/n$ chance of being correct. Can a sequential guessing strategy do better?

Suppose that the permuted numbers are exposed sequentially, one at a time, π_1 first, then π_2 , and so on. Before the next number is exposed the gambler is given the opportunity of guessing whether the next number is 1 or passing on a guess at that turn. The process halts on the turn after a gambler's guess is checked or if the gambler passes and the number 1 is exposed. A sequential, randomised guessing strategy may be described as follows. The gambler initially bets with probability p_1 that the first number π_1 is 1 and with probability $1 - p_1$ passes on making an initial guess. If he passes on the first turn then, after π_1 is exposed, if $\pi_1 \neq 1$, the gambler, with probability p_2 , elects to guess that π_2 is 1 and passes on the next turn with probability $1 - p_2$. Proceeding in this fashion, if the gambler has passed on the first $k - 1$ turns and the numbers π_1 through π_{k-1} are none of them equal to 1, then with probability p_k he guesses that π_k is 1 and passes on the guess with probability $1 - p_k$. If the first $n - 1$ permuted numbers have been exposed without either a guess being made or 1 being encountered then, manifestly, π_n must be 1 and accordingly he guesses with probability $p_n = 1$ that π_n is 1. The selection of the numbers p_1, \dots, p_{n-1} with $p_n = 1$ determines the sequential, randomised guessing strategy employed by the gambler. What is the probability that a given strategy makes a successful guess?

Let Q_n be the probability that a given sequential, randomised strategy produces a successful guess for the location of a given number out of n permuted numbers. While there is a combinatorial explosion if we attempt to look at the problem as a whole, the picture simplifies enormously if we consider it one turn at a time. Conditioned on a guess being made at the first turn, the probability that it is successful is clearly $1/n$. On the other hand, if a guess is not made at the first turn, the sequence will continue if, and only if, $\pi_1 \neq 1$, an

³W. A. Whitworth, "Arrangements of m things of one sort and n things of another sort under certain conditions of priority", *Messenger of Math.*, vol. 8, pp. 105–114, 1878.

event of probability $1 - 1/n$ unaffected by the gambler's decision on whether or not to guess at the first turn. But then we have a permuted set of $n - 1$ numbers remaining, one of which is 1, and the probability that the sequential strategy under consideration will successfully unearth the location of 1 in this reduced set is, by definition, Q_{n-1} . Accordingly, by conditioning on whether the strategy elects to guess at the first turn or not, we obtain the recurrence

$$Q_n = p_1 \frac{1}{n} + (1 - p_1)(1 - \frac{1}{n})Q_{n-1} \quad (n \geq 2).$$

The nature of the recurrence becomes clearer if we multiply both sides by n . Setting $T_k := kQ_k$ for each k , we then obtain

$$T_n = p_1 + (1 - p_1)T_{n-1} \quad (n \geq 2)$$

with the obvious boundary condition $T_1 = Q_1 = 1$. Then it is trivial to verify that $T_2 = p_1 + (1 - p_1)T_1 = 1$ and $T_3 = p_1 + (1 - p_1)T_2 = 1$, leading to the delicious conjecture that $T_n = 1$ for all n . Verification is a simple matter of induction. It follows that $Q_n = 1/n$ for any sequential, randomised guessing strategy. Thus, *sequential guessing yields no improvement over the a priori guess.* ►

The theorem of total probability is simply a mathematical codification of the dictum: *the whole is equal to the sum of its parts*. The result should be viewed as just a restatement of the countable additivity property of probability measure. Typical examples of its utility are sketched in the following sections; the contexts, while classical, continue to be relevant in the modern theory.

4 Le problème des rencontres, matchings

A group of n increasingly inebriated sailors on shore leave is making its unsteady way from pub⁴ to pub. Each time the sailors enter a pub they (being very polite, Gilbertian sailors, even in their inebriated state) take off their hats and leave them at the door. On departing for the next pub, each intoxicated sailor picks up one of the hats at random. What is the probability Q_n that no sailor retrieves his own hat?

The setting is somewhat dated—who wears hats outside of Ascot any more?—and sounds like a party game, of no great moment; but it turns out to be surprisingly ubiquitous. Conditioning quickly leads to the answer but the argument is subtle and warrants scrutiny.

Write M_{ij} for the event that sailor i retrieves (is matched with) hat j and let A_n be the event that no sailor is matched with his own hat. If no sailor is to retrieve his hat then, *a fortiori*, the n th sailor must be in possession of someone

⁴Short for PUBLIC HOUSE; a tavern or hostelry where intoxicating drinks may be purchased.

Conditional Probability

else's hat. By total probability,

$$Q_n = P(A_n) = \sum_{j=1}^n P(A_n | M_{nj}) P(M_{nj}) = \frac{1}{n} \sum_{j=1}^{n-1} P(A_n | M_{nj}) \quad (4.1)$$

as the match M_{nj} occurs with probability $1/n$ for each j and the event A_n cannot occur if sailor n is matched with his own hat. Conditioned on the event that sailor n is matched with hat j , the event A_n can occur in two distinct fashions depending on the matching of sailor j : either sailor j is matched with hat n or sailor j is matched with a different hat.

In the first case, sailor j has $n - 1$ possible matches of which hat n is one and so the conditional match of sailor j to hat n has probability $1/(n - 1)$. Once sailor j has been matched to hat n , the remaining $n - 2$ sailors have their own hats available for matching as sailors j and n and their corresponding hats have been removed from the pool. Thus, conditioned upon the twin matches M_{nj} and M_{jn} , the probability of the event A_n is just Q_{n-2} as the remaining $n - 2$ sailors must each avoid their own hats if A_n is to occur. Thus, conditioned upon the match M_{nj} , the probability that A_n occurs with sailor j matched to hat n is $\frac{1}{n-1} Q_{n-2}$.

Alternatively, with sailor n and hat j removed from the pool, sailor j may be matched to a hat other than n . The situation is now reminiscent of a matching problem for $n - 1$ sailors and a little labelling legerdemain makes it clear that we are, in fact, dealing exactly with that setting. Each of the $n - 2$ sailors other than j has a hat which we may label with his name; as the n th sailor has left his hat behind, we may temporarily label his hat with sailor j 's name. We now have $n - 1$ sailors, each with a labelled hat. The probability that j is not matched to "his" hat, and none of the other sailors are matched to theirs is, by definition, Q_{n-1} . And so, conditioned upon match M_{nj} , the probability that A_n occurs with sailor j matched to a hat other than n is just Q_{n-1} .

Thus, by conditioning additionally upon whether sailor j is matched to hat n or not, we obtain (additivity, again!)

$$P(A_n | M_{nj}) = \frac{1}{n-1} Q_{n-2} + Q_{n-1} \quad (j \neq n),$$

the right-hand side being independent of j , as indeed it must by reasons of symmetry. Substitution into (4.1) shows that

$$Q_n = \frac{1}{n} Q_{n-2} + \frac{n-1}{n} Q_{n-1}, \quad \text{or, rewritten,} \quad Q_n - Q_{n-1} = -\frac{1}{n} [Q_{n-1} - Q_{n-2}].$$

Introducing the difference operator $\Delta Q_i := Q_i - Q_{i-1}$, we may write the expression compactly in the form

$$\Delta Q_n = -\frac{1}{n} \Delta Q_{n-1}.$$

While the recurrence holds formally for $n \geq 2$, it will be convenient to extend its domain of validity to all $n \geq 1$ with the natural choices $Q_0 := 1$ and $\Delta Q_0 := 1$.

Then the recurrence yields $\Delta Q_1 = -1$ and $\Delta Q_2 = 1/2$, which is consistent with the probabilities $Q_1 = 0$ and $Q_2 = 1/2$ as may be written down by inspection. The general solution beckons and we may write

$$\Delta Q_n = \left(-\frac{1}{n}\right)\left(-\frac{1}{n-1}\right) \cdots \left(-\frac{1}{2}\right)\left(-\frac{1}{1}\right)\Delta Q_0 = \frac{(-1)^n}{n!}$$

for each $n \geq 1$. As $Q_n = Q_{n-1} + \Delta Q_n$, an easy induction now shows that

$$Q_n = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n \frac{1}{n!} = \sum_{k=0}^n \frac{(-1)^k}{k!} \quad (n \geq 0).$$

We recognise in the sum on the right-hand side a truncated exponential series:

$$e^{-1} = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}.$$

The series converges very rapidly and for even moderate values of n we obtain the approximation $Q_n \approx e^{-1} = 0.3678 \dots$. For a party of five sailors the correct answer is $Q_5 = 0.3666 \dots$ so that the exponential series approximation is already good to within about one part in a thousand. Thus, *the probability that at least one sailor retrieves his own hat is approximately two-thirds and is almost independent of the number of sailors in the party*.

The problem is more usually expressed in terms of *matchings*. If n letters are randomly matched with n addressed envelopes, the probability that there is a matching of at least one letter with the proper envelope is $1 - Q_n$. Likewise, the probability that in a random shuffle of a standard 52-card deck at least one card returns to its original position is again $1 - Q_n$, where n equals 52, and the probability of at least one matching is again approximately two-thirds.

In the classical literature the problem is known as *le problème des rencontres*. De Montmort posed it in 1708 and discovered its surprising solution by direct combinatorial methods. We shall return to it in Section IV.1 with a classical approach which follows in de Montmort's footsteps. The problem repays study; it continues to be informative and crops up frequently in applications.

5 Pólya's urn scheme, spread of contagion

A town has a population of $r + s$ individuals belonging to two rival sects: the royalists are initially r in number with the seditious numbering s . At each epoch a group of a new arrivals makes its way to the town and casts its lot with one sect or the other depending on the allegiance of the first inhabitant it meets. What can be said about the ebb and flow of the proportions of royalists and seditious in the town?

The setting, shorn of political colour, is that of *Pólya's urn scheme*. An urn initially contains r red balls and s black balls. A ball is selected at random *but not removed* and a balls of the same colour as the selection are added to the urn. The process is then repeated with a balls of one colour or the other added to the urn at each epoch. With each addition the population of the urn increases by a and it is helpful to imagine that the urn has unbounded capacity. For the model it is not necessary, however, that a is a positive integer; we may take a as having any integral value. If $a = 0$ then the status quo in the urn is preserved from epoch to epoch. If a is negative, however, the population of the urn *decreases* with each step and, assuming both r and s are divisible by $|a|$, terminates with first one subpopulation being eliminated, then the other.

To obviate trivialities, suppose now that $a \neq 0$. The probability of selecting a red ball initially is $r/(r+s)$ and in this case the red population increases to $r+a$ after one epoch; contrariwise, a black ball is selected with probability $s/(r+s)$ in which case the black population increases to $s+a$ after one epoch. (It should be clear that in this context words like "add" and "increase" should be interpreted depending on the sign of a ; if a is negative they correspond to "subtract" and "decrease", respectively.)

The probability of selecting two reds in succession is obtained by a simple conditional argument with an obvious notation as

$$P(R_1 \cap R_2) = P(R_2 | R_1) P(R_1) = \frac{r+a}{r+s+a} \cdot \frac{r}{r+s},$$

so that the probability that the urn contains $r+2a$ red balls after two epochs is $r(r+a)/((r+s)(r+s+a))$. Arguing likewise, the probability of selecting two blacks in succession is given by

$$P(B_1 \cap B_2) = \frac{s+a}{r+s+a} \cdot \frac{s}{r+s}$$

so that, after two epochs, the probability that there are $s+2a$ black balls in the urn is $s(s+a)/((r+s)(r+s+a))$. As one red ball and one black ball may be selected in the first two epochs with red first and black next or vice versa, by total probability the chance of observing $r+a$ red balls and $s+a$ black balls in the urn after two epochs is given by entirely similar conditioning considerations to be

$$\frac{s}{r+s+a} \cdot \frac{r}{r+s} + \frac{r}{r+s+a} \cdot \frac{s}{r+s} = \frac{2rs}{(r+s)(r+s+a)}.$$

The argument readily generalises. Let $P_n(i)$ represent the probability that the urn contains i red balls after n epochs. It is clear that $P_n(i)$ can be non-zero only for $i = r+ka$ where the integer k can vary over at most $0 \leq k \leq n$. (If a is negative the experiment can terminate earlier due to an exhaustion of balls in the urn.) Now, after n epochs, the urn can contain $r+ka$ red balls [and, consequently, also $s+(n-k)a$ black balls] if, and only if, k red balls and $n-k$

black balls were selected. Each selection of red increases the red population by a ; each selection of black increases the black population by a . Each ordering of k red selections and $n - k$ black selections hence has the same probability and it follows that

$$P_n(r + ka) = \binom{n}{k} \frac{r(r+a) \cdots (r+(k-1)a) \cdot s(s+a) \cdots (s+(n-k-1)a)}{(r+s)(r+s+a) \cdots (r+s+(n-1)a)}. \quad (5.1)$$

Generalised binomial notation allows us to compact the equation into a more intuitive and elegant form.

For any real x and positive integer k , we write $x^k := x(x-1) \cdots (x-k+1)$ in the “falling factorial” notation for a product of k terms starting at x with each successive term differing from the previous one in one unit. The generalised binomial coefficients are now defined by $\binom{x}{k} := x^k/k!$ (see Problems I.1–5). As a simple consequence, we may relate “negative” binomial coefficients to the ordinary kind via

$$\binom{-x}{k} = \frac{(-x)(-x-1) \cdots (-x-k+1)}{k!} = (-1)^k \binom{x+k-1}{k}.$$

With notation for the generalised binomial in hand we return to the occupancy probabilities $P_n(\cdot)$. Factoring out a from each of the product terms on the right of (5.1), the fraction on the right may be written more compactly in the form

$$\begin{aligned} & \frac{\frac{r}{a} \left(\frac{r}{a} + 1 \right) \cdots \left(\frac{r}{a} + k - 1 \right) \cdot \frac{s}{a} \left(\frac{s}{a} + 1 \right) \cdots \left(\frac{s}{a} + n - k - 1 \right)}{\frac{r+s}{a} \left(\frac{r+s}{a} + 1 \right) \cdots \left(\frac{r+s}{a} + n - 1 \right)} \\ &= \frac{\left(\frac{r}{a} + k - 1 \right)^k \left(\frac{s}{a} + n - k - 1 \right)^{n-k}}{\left(\frac{r+s}{a} + n - 1 \right)^n}. \end{aligned}$$

Grouping terms on the right of (5.1) we hence obtain

$$P_n(r + ka) = \frac{\binom{\frac{r}{a} + k - 1}{k} \binom{\frac{s}{a} + n - k - 1}{n - k}}{\binom{\frac{r+s}{a} + n - 1}{n}} = \frac{\binom{-r/a}{k} \binom{-s/a}{n-k}}{\binom{-(r+s)/a}{n}}.$$

For the particular case when $a = 1$ and there is one new addition to the population each epoch, we obtain the simple expression

$$P_n(r + k) = \frac{\binom{-r}{k} \binom{-s}{n-k}}{\binom{-(r+s)}{n}} \quad (0 \leq k \leq n).$$

If $a = -1$, on the other hand, we have sampling without replacement and the population decreases by one at each epoch down to zero when $n = r + s$. For this case substitution shows that

$$P_n(r - k) = \frac{\binom{r}{k} \binom{s}{n-k}}{\binom{r+s}{n}} \quad (n - s \leq k \leq r),$$

the extremes of the range for k indicating where one political faction or the other is exterminated. This is the *hypergeometric distribution*; we will see it arising less serendipitously in Section VIII.10.

On an apolitical note, the setting of Pólya's urn scheme may be taken as a crude representation of the spread of contagion. The more individuals who are infected in the population, the greater is the chance new arrivals become infected.

6 The Ehrenfest model of diffusion

Two chambers labelled, say, left and right, and separated by a permeable membrane contain N indistinguishable particles distributed between them. At each epoch a randomly selected particle exchanges chambers and passes through the membrane to the other chamber, the more populous chamber being more likely to proffer the exchange particle. The setting that has been described provides a primitive model of a diffusion or osmotic process and was proposed by P. and T. Ehrenfest.⁵ The model has been expanded widely and continues to be important in the theory of diffusion processes.

The *state* of the system at any epoch may be described by the number of particles in, say, the left chamber. The state then is an integer between 0 and N . If the current state is k then, at the next transition epoch, the state becomes either $k - 1$ or $k + 1$ depending on whether the left chamber proffers the exchange particle or receives it. It is natural in this setting to assume that the chance of the exchange particle being drawn from a given chamber is proportional to its current population. Thus, a particle is more likely to be lost by the more populous chamber so that there is a stochastic restoring force towards population equilibrium in the two chambers. This picture is consonant with the kind of stochastic diffusion or osmosis from regions of higher concentration to regions of lesser concentration that one sees in a variety of natural situations.

To keep with later terminology and notation, let p_{jk} denote the conditional probability that at a given transition epoch the system state changes from j to k . The quantities p_{jk} are naturally called *transition probabilities* for the system. In the Ehrenfest model transitions are only possible to neighbouring states with transition probabilities proportional to the population of the chamber that offers the exchange particle; accordingly, we must have $p_{k,k-1} = k/N$ and $p_{k,k+1} = (N - k)/N$ for $0 \leq k \leq N$. Here, as in the previous examples, it is most natural to specify the underlying probability measure in terms of conditional probabilities.

Suppose that initially the system is in state k with probability $u_k^{(0)}$. Starting with a state selected randomly according to the system of probabili-

⁵P. Ehrenfest and T. Ehrenfest, "Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem", *Physikalische Zeitschrift*, vol. 8, pp. 311–314, 1907.

ties (or *probability distribution*) $\{u_0^{(0)}, \dots, u_N^{(0)}\}$ the system progresses through a stochastic succession of states governed by the state transition probabilities p_{jk} at each step. Suppose that after n epochs the probability that the system is in state k is denoted $u_k^{(n)}$. Then the collection of probabilities $\{u_0^{(n)}, \dots, u_N^{(n)}\}$ determines the distribution of states after n epochs. It is natural now to query whether the system reaches a stochastic equilibrium after a large number of steps when, presumably, initial transients caused by the particular choice of initiating distribution $\{u_k^{(0)}\}$ have died away.

The first state transition provides a model for subsequent steps. The probability $u_k^{(1)}$ that the system enters state k after one step may be determined by conditioning on the starting initial state. Total probability mops up and we obtain

$$u_k^{(1)} = \sum_{j=0}^N u_j^{(0)} p_{jk} = u_{k-1}^{(0)} \cdot \frac{N-k+1}{N} + u_{k+1}^{(0)} \cdot \frac{k+1}{N} \quad (0 \leq k \leq N)$$

as the initial state has to be either $k-1$ or $k+1$ if the next state is to be k . The initial distribution of states $\{u_k^{(0)}\}$ is hence replaced by the distribution of states $\{u_k^{(1)}\}$ after one step. The process now repeats. Given the distribution of states $\{u_k^{(n-1)}\}$ after $n-1$ steps, the distribution of states $\{u_k^{(n)}\}$ after the n th transition is given by

$$u_k^{(n)} = \sum_{j=0}^N u_j^{(n-1)} p_{jk} = u_{k-1}^{(n-1)} \cdot \frac{N-k+1}{N} + u_{k+1}^{(n-1)} \cdot \frac{k+1}{N} \quad (6.1)$$

for each $0 \leq k \leq N$. Starting with the initial distribution of states, we may recursively determine the evolution of state distribution, one step at a time.

If initial transients are to be suppressed and the system settle down to a stochastic equilibrium then there must exist a system of state probabilities $\{u_0, \dots, u_N\}$ with $u_k^{(n)} \rightarrow u_k$ for each k as $n \rightarrow \infty$. If that is the case then it is clear from (6.1) that we must have

$$u_k = u_{k-1} \cdot \frac{N-k+1}{N} + u_{k+1} \cdot \frac{k+1}{N} \quad (6.2)$$

for each k . Any system of probabilities $\{u_0, \dots, u_N\}$ satisfying (6.2) is invariant in that if the state probabilities at any epoch are given by $\{u_k\}$ then they remain unaltered thereafter. Any such system of probabilities is hence said to form an *invariant* (or *stationary*) distribution.

We may proceed inductively to determine the values u_k one at a time from the governing equations (6.2). When $k=0$ we obtain $u_0 = u_1/N$ or $u_1 = Nu_0$. Setting $k=1$ and 2 in turn, we find then that

$$\begin{aligned} u_1 &= u_0 + u_2 \cdot \frac{2}{N}, \quad \text{or, equivalently, } u_2 = \frac{N}{2}[u_1 - u_0] = u_0 \cdot \frac{N(N-1)}{2!}, \\ u_2 &= u_1 \cdot \frac{N-1}{N} + u_3 \cdot \frac{3}{N}, \quad \text{or } u_3 = \frac{N}{3}[u_2 - u_1 \cdot \frac{N-1}{N}] = u_0 \cdot \frac{N(N-1)(N-2)}{3!}. \end{aligned}$$

Conditional Probability

The general form of the solution $u_k = u_0 \binom{N}{k}$ is now apparent and is easily verified by induction as Pascal's triangle yields

$$\binom{N}{k-1} \frac{N-k+1}{N} + \binom{N}{k+1} \frac{k+1}{N} = \binom{N-1}{k-1} + \binom{N-1}{k} = \binom{N}{k}.$$

As $\{u_0, \dots, u_N\}$ forms a system of probabilities, the normalisation condition

$$1 = \sum_{k=0}^N u_k = u_0 \sum_{k=0}^N \binom{N}{k} = u_0 2^N$$

shows that $u_0 = 2^{-N}$ and it follows that

$$u_k = 2^{-N} \binom{N}{k} \quad (0 \leq k \leq N) \quad (6.3)$$

is the unique stationary distribution for the Ehrenfest model of diffusion.

The form of the solution is of interest in its own right: this is the important *binomial distribution* and we will return to it in more depth in subsequent chapters. Two observations regarding its nature will suffice for our purposes here. Observe first that as $\binom{N}{k} = \binom{N}{N-k}$ the solution for u_k is symmetric around $k = N/2$ as we would anticipate on grounds of symmetry for the occupancies of the two chambers. Furthermore, it is easy to verify that $\binom{N}{k}/\binom{N}{k-1} \geq 1$ if, and only if, $k \leq (N+1)/2$. Consequently, u_k increases monotonically in the range $0 \leq k \leq N/2$ and decreases monotonically thereafter. In particular, an equidistribution of particles between the two chambers is probabilistically the most favoured eventuality. Deviations from equidistribution see a "restoring force" towards parity. For another diffusion model along similar lines the reader is directed to Problem 22.

The careful reader may cavil that while we have shown that a unique stationary distribution $\{u_k\}$ exists for the Ehrenfest model it does not necessarily follow that the probabilities $u_k^{(n)}$ converge pointwise to u_k for every choice of starting distribution of states. The result that we seek is actually a particular illustration of a very general result on distributional convergence that holds under a wide range of circumstances but we will not pursue it further here.



EXTENSION: BIRTH-DEATH CHAINS

The Ehrenfest model of diffusion is illustrative of a system which at any given epoch is in one of a countable collection of states which, for purposes of definiteness, we may identify with the positive integers $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$. At each succeeding epoch the system may probabilistically change its state *but only to a neighbouring state*. Write p_{jk} for the probability that the system makes a transition to state k conditioned on the event that it was in state j immediately prior to the transition. The requirement that transitions may only be made to neighbouring states is captured by the condition $p_{jk} = 0$ if $|j - k| > 1$

or, equivalently, $p_{k,k-1} + p_{kk} + p_{k,k+1} = 1$ for each k . Implicit in this structure is the understanding that transitions are governed solely by the current state and, in particular, do not depend upon the past history of state transitions. The system then stochastically meanders through a chain of states X_0, X_1, X_2, \dots where, for each n , X_n denotes the state of the system at epoch n and $|X_{n+1} - X_n| \leq 1$.

Some colourful terminology has evolved to describe chains of this nature. Unit increments in state are identified with births and, correspondingly, unit decrements in state are identified with deaths. Random sequences of this form are hence called *birth-death chains*.

Let $u_k^{(n)}$ denote the probability that the system is in state k at epoch n . By conditioning on possible states the system could be in at epoch $n-1$ we obtain the recurrence

$$u_k^{(n)} = u_{k-1}^{(n-1)} p_{k-1,k} + u_k^{(n-1)} p_{kk} + u_{k+1}^{(n-1)} p_{k+1,k} \quad (k \geq 0). \quad (6.4)$$

If a stationary distribution of states $\{u_k\}$ is to exist then we must have perforce

$$u_k = u_{k-1} p_{k-1,k} + u_k p_{kk} + u_{k+1} p_{k+1,k}$$

for each k . Beginning with $k = 0$ and proceeding sequentially (bearing in mind that $p_{kk} = 1 - p_{k,k-1} - p_{k,k+1}$) we obtain the sequence of putative solutions

$$u_1 = u_0 \frac{p_{01}}{p_{10}}, \quad u_2 = u_0 \frac{p_{01}p_{12}}{p_{21}p_{10}}, \quad u_3 = u_0 \frac{p_{01}p_{12}p_{23}}{p_{32}p_{21}p_{10}},$$

and so on. As we must have $\sum_{k \geq 0} u_k = 1$ if $\{u_k\}$ is to be an honest probability distribution, it follows that *a stationary distribution exists for the birth-death chain if, and only if,*

$$\sum_{k=0}^{\infty} \frac{p_{01}p_{12} \cdots p_{k-2,k-1}p_{k-1,k}}{p_{k,k-1}p_{k-1,k-2} \cdots p_{21}p_{10}} < \infty;$$

under this condition there exists a unique stationary distribution $\{u_k\}$ for the chain with invariant state probabilities given by

$$u_k = u_0 \frac{p_{01}p_{12} \cdots p_{k-2,k-1}p_{k-1,k}}{p_{k,k-1}p_{k-1,k-2} \cdots p_{21}p_{10}} \quad (k \geq 1) \quad (6.5)$$

and where u_0 is determined by the normalisation condition to be

$$u_0 = \left[\sum_{k=0}^{\infty} \frac{p_{01}p_{12} \cdots p_{k-2,k-1}p_{k-1,k}}{p_{k,k-1}p_{k-1,k-2} \cdots p_{21}p_{10}} \right]^{-1}.$$

Diffusion processes are natural candidates for birth-death models. As we have seen, we may identify the Ehrenfest model of diffusion as a birth-death chain on a finite number of states $\{0, 1, \dots, N\}$ and with transition probabilities $p_{k,k-1} = k/N$ and $p_{k,k+1} = (N-k)/N$ for $0 \leq k \leq N$. The stationary probabilities in this case are given by (6.3). Queuing processes form another class of natural candidates. Application domains include the classical consideration of call arrivals to a telephone switching system, ticket sales ebbing and flowing at an eagerly anticipated event, the arrival and departure of cars at a metered highway access ramp, packet arrivals at a busy internet node, and the buffering of real-time video or audio data streams in a computer.

7 Bayes's rule for events, the MAP principle

Set intersection is a commutative operation (that is to say, it is insensitive to order). Consequently,

$$\mathbf{P}(A | H) \mathbf{P}(H) = \mathbf{P}(A \cap H) = \mathbf{P}(H | A) \mathbf{P}(A),$$

where the usage of conditional probabilities implicitly requires that A and H are events of non-zero probability. It follows that

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

which is little more than a restatement of the definition of conditional probability. Coupling the “reversed” conditional probability with total probability we obtain the useful theorem of Bayes. To obviate the necessity for frequent caveats on the existence of conditional probabilities the reader should assume that all events that are being conditioned on have strictly positive probability.

BAYES'S RULE FOR EVENTS *Suppose $\{A_j, j \geq 1\}$ is a measurable partition of Ω and H is any event. Then, for each k , we have*

$$\mathbf{P}(A_k | H) = \frac{\mathbf{P}(H | A_k) \mathbf{P}(A_k)}{\sum_{j \geq 1} \mathbf{P}(H | A_j) \mathbf{P}(A_j)}.$$

Bayes's rule finds application in a variety of settings where it is most natural to specify the conditional probabilities $\mathbf{P}(H | A_k)$ in the problem statement but it is the “reversed” conditional probabilities $\mathbf{P}(A_k | H)$ that are desired. Applications are explored in this and the following sections.

EXAMPLES: 1) *Dice.* Two dice are thrown. Given that the sum of the face values is 9, what is the probability that a 6 was thrown? While we could work directly with the sample space in this simple setting it is instructive to lay out the underlying conditioning mechanism.

Write H for the event that the sum of face values is 9, A_j for the event that the first die shows j , B_k for the event that the second die shows k . The sample points of the experiment are the 36 equally likely pairs (j, k) with $1 \leq j, k \leq 6$. As $A_j \cap B_k = \{(j, k)\}$, it follows that $\mathbf{P}(A_j \cap B_k) = \mathbf{P}\{(j, k)\} = 1/36$ and so additivity yields $\mathbf{P}(A_j) = \sum_k \mathbf{P}(A_j \cap B_k) = 1/6$ for each j ; likewise, $\mathbf{P}(B_k) = \sum_j \mathbf{P}(A_j \cap B_k) = 1/6$ for each k . The events A_1, \dots, A_6 partition the sample space and so, by total probability, $\mathbf{P}(H) = \sum_j \mathbf{P}(H | A_j) \mathbf{P}(A_j)$. Now $\mathbf{P}(H | A_j) = \mathbf{P}(B_{9-j}) = 1/6$ when $3 \leq j \leq 6$, these being the only cases contributing to the probability of H as for H to occur both dice have to show at least 3. Thus, $\mathbf{P}(H) = \sum_{j=3}^6 \mathbf{P}(H | A_j) \mathbf{P}(A_j) = 4/36$. Bayes's rule now shows quickly that the conditional probability that the first die shows a 6 given that the sum of the dice

is 9 is given by $\mathbf{P}(A_6 \mid H) = \frac{1/36}{4/36} = 1/4$. An entirely similar argument shows likewise that the conditional probability that the second die shows a 6 given that the sum is 9 is also $\mathbf{P}(B_6 \mid H) = 1/4$. As the events A_6 and B_6 are mutually exclusive *conditioned on H* it follows that the conditional probability that a 6 was thrown given that the sum of face values is 9 is $1/4 + 1/4 = 1/2$.

2) *Return to families.* A telephone survey of a family with two children is answered by a boy. What is the probability the other child is also a boy?

For $j, k \in \{1, 2\}$, let B_j be the event that the j th child is a boy and let G_k be the event that the k th child is a girl, with the natural, uncontroversial assignment of probabilities

$$\mathbf{P}(B_1 \cap B_2) = \mathbf{P}(B_1 \cap G_2) = \mathbf{P}(G_1 \cap B_2) = \mathbf{P}(G_1 \cap G_2) = 1/4. \quad (7.1)$$

Let B denote the event that a boy answers the phone. Given that a boy answers the phone, as the other child can be a boy if, and only if, both children are boys, the event of interest is $B_1 \cap B_2$. We may now write down an expression for the desired conditional probability via Bayes's rule,

$$\mathbf{P}(B_1 \cap B_2 \mid B) = \frac{\mathbf{P}(B \mid B_1 \cap B_2) \mathbf{P}(B_1 \cap B_2)}{\mathbf{P}(B)} = \frac{1/4}{\mathbf{P}(B)}$$

as it is certain that a boy answers the phone if both children are boys. It only remains now to evaluate the probability that a boy answers the phone. Total probability yields

$$\begin{aligned} \mathbf{P}(B) &= \mathbf{P}(B \mid B_1 \cap B_2) \mathbf{P}(B_1 \cap B_2) + \mathbf{P}(B \mid B_1 \cap G_2) \mathbf{P}(B_1 \cap G_2) \\ &\quad + \mathbf{P}(B \mid G_1 \cap B_2) \mathbf{P}(G_1 \cap B_2) + \mathbf{P}(B \mid G_1 \cap G_2) \mathbf{P}(G_1 \cap G_2) \\ &= \frac{1}{4} [1 + \mathbf{P}(B \mid B_1 \cap G_2) + \mathbf{P}(B \mid G_1 \cap B_2)]. \end{aligned}$$

We appear to be at an impasse as the natural assumption that the genders are equally likely of occurrence does not give guidance as to how we should allocate numbers for the conditional probabilities on the right. Further consideration shows that the sample space could use some edification. While at first blush the sample space for this problem appears to coincide with that of Example 1.3, a little thought suggests a difference in the setting: birth order and sex distribution could conceivably affect telephone answering mores. Building on this idea it appears productive to expand the sample space to include not only the sexes of the children by birth order but also whether the elder or the younger answers the phone though it is not clear at this stage how to allocate probabilities to these outcomes—all that can be concluded is that any allocation of probabilities should be consistent with the equally likely assumption on sexes. The expanded sample space with consistent probabilities is then of the form given in Table 4 where each sample point lists the sexes of the children

Conditional Probability

Outcomes	bb1	bb2	bg1	bg2	gb1	gb2	gg1	gg2
Probabilities	$\frac{1}{4}\alpha$	$\frac{1}{4}(1-\alpha)$	$\frac{1}{4}\beta$	$\frac{1}{4}(1-\beta)$	$\frac{1}{4}\gamma$	$\frac{1}{4}(1-\gamma)$	$\frac{1}{4}\delta$	$\frac{1}{4}(1-\delta)$

Table 4: An expanded sample space for Example 2.

by birth order followed by an identification of whether the first or the second child answers the phone and the parameters α , β , γ , and δ take arbitrary values in the unit interval. It is easy to verify now that the uniform gender-order probabilities (7.1) hold for any choice of these parameters.

With the sample space clearly identified it is now easy to write down the conditional probabilities

$$\mathbf{P}(B | B_1 \cap G_2) = \frac{1}{4}\beta / \frac{1}{4} = \beta, \quad \mathbf{P}(B | G_1 \cap B_2) = \frac{1}{4}(1-\gamma) / \frac{1}{4} = 1-\gamma,$$

whence $\mathbf{P}(B) = \frac{1}{4}(2 + \beta - \gamma)$, and so

$$\mathbf{P}(B_1 \cap B_2 | B) = \frac{\frac{1}{4}}{\frac{1}{4}(2 + \beta - \gamma)} = \frac{1}{2 + \beta - \gamma}.$$

There are various more or less “natural” choices for the parameters β and γ . For instance, if $\beta = \gamma$ (including the special case $\alpha = \beta = \gamma = \delta = 1/2$ when either child is equally likely to answer the phone), we obtain $\mathbf{P}(B_1 \cap B_2 | B) = 1/2$. This is the “naïve” independence argument that we rejected under the assumptions of Example 1.3 but, as we now see, is a valid possibility in this modified setting. Another possibility is when $\beta = 1$ and $\gamma = 0$ corresponding to the case when a boy, if there is one, will always answer the phone. In this case $\mathbf{P}(B_1 \cap B_2 | B) = 1/3$; this is exactly the situation analysed in Example 1.3. Still another possibility arises when $\beta = 0$ and $\gamma = 1$ corresponding to the case when a girl, if there is one, will always answer the phone. In this case $\mathbf{P}(B_1 \cap B_2 | B) = 1$. As we see, in general, $\mathbf{P}(B_1 \cap B_2 | B)$ may take any value between $1/3$ and 1 depending on the assumptions made. ►

In applications one frequently encounters a natural notion of causality or “time’s arrow” where the mutually exclusive events $\{A_k\}$ take on the rôle of priors—initial settings for the problem—while the event H constitutes an “observable” or result of the experiment. An interpretation along these lines leads to a natural question: *given that H is observed, which A_k engendered it?* In these settings one typically has access to the conditional probabilities $\mathbf{P}(H | A_k)$ and the goal is to infer the *a posteriori* probabilities $\mathbf{P}(A_k | H)$ (that is to say, probabilities determined after the fact on the basis of empirical evidence). If the *a priori* probabilities $\mathbf{P}(A_k)$ (that is to say, probabilities based upon theoretical deduction prior to performing an experiment) are known then Bayes’s rule points

the way: given that H has occurred, the maximum *a posteriori* probability (or MAP) principle selects that A_k for which the *a posteriori* probability $P(A_k | H)$ is the largest.

This kind of approach is unexceptionable when there is no controversy over the assignment of the *a priori* probabilities $P(A_k)$. But, as we saw in the previous example, it is not always clear what the assignment of these *a priori* probabilities should be. In such cases, the Bayesian proceeds by making an assumption about the nature of the prior much to the horror of the frequentist who insists upon empirical evidence. The validity of the Bayesian choice may be defended on metaphysical but not mathematical grounds. Its success ultimately depends on whether the assumptions succeed in adequately capturing the physical reality.

8 Laplace's law of succession

The following urn problem is historical and illustrative. A sequence of $N + 1$ urns is given, numbered for convenience from 0 through N . Each urn has N balls with urn k containing k red balls and $N - k$ black balls. An urn is first picked at random; balls are then successively picked at random from the chosen urn with each ball replaced before the next ball is picked. If in a sequence of m drawings, m red balls appear in succession, what is the probability that the $(m + 1)$ th draw also results in a red ball?

Suppose one were to (temporarily) suspend belief in physical laws and assume (a) that our universe was randomly selected from a large number of possible universes, and (b) that the daily rising and setting of the sun is governed not by predictable physical laws but by a repeated series of random experiments. This model maps neatly into our urn problem: the universes are represented by the urns; once a universe (i.e., urn) is selected our cosmic gambler randomly selects a ball (with replacement—balls, if not energy, are conserved in this universe) each day; the sun rises on a given day if the ball drawn on that day is red. Our question may be posed in a somewhat more memorable fashion in this model: *what is the probability that the sun will rise tomorrow given that it has risen for m consecutive days (since creation) without break?*

Write A_m for the event that m successive balls drawn were red and U_k for the event that urn k is chosen. Then the probability of the event U_k is $1/(N + 1)$ while, conditioned on having chosen urn k , the probability of A_m is exactly k^m/N^m . The probabilistic assumptions here bear a little investigation.

What is the sample space under consideration? After some thought we may settle on a sample space corresponding to the selection of an urn followed by an unending sequence of ball selections with replacement. A sample point of the experiment may be represented in the form of a sequence $\omega = (u; v_1, v_2, v_3, \dots, v_m, v_{m+1}, \dots)$ where u takes values in $\{0, 1, \dots, N\}$ and each v_i takes values in $\{0, 1\}$, red and black represented by 1 and 0, respectively. Our

Conditional Probability

earlier investigations into continuous sample spaces show that this set of sample points has the cardinality of the continuum. The event A_m in this space corresponds to the set of sample points for which $v_1 = v_2 = \dots = v_m = 1$ while the event U_k corresponds to the set of sample points for which $u = k$. It is natural in this setting to assign the *a priori* probabilities

$$P(U_k) = \frac{1}{N+1} \text{ and } P(A_m | U_k) = \left(\frac{k}{N}\right)^m. \quad (8.1)$$

The event $U_k \cap A_m$ specifies a *cylinder set* $C_{k,m}$ consisting of those sample points $\omega = (u; v_1, v_2, v_3, \dots, v_m, v_{m+1}, \dots)$ for which $u = k$ and $v_1 = v_2 = \dots = v_m = 1$ (with v_{m+1}, v_{m+2}, \dots arbitrarily specified). (The term “cylinder” is in analogy with the usual geometric notion of a three-dimensional cylinder with one coordinate varying freely.) Thus, our assignment of probabilities merely says that

$$P(C_{k,m}) = P(U_k \cap A_m) = P(A_m | U_k) P(U_k) = \frac{1}{N+1} \left(\frac{k}{N}\right)^m.$$

The same process of specification can be used to systematically specify the probabilities of any other cylinder set. Thus, if $C_{k,r,m}$ specifies a cylinder set consisting of sample points corresponding to the selection of the k th urn and where the first m draws result in a specific arrangement of r red balls and $m - r$ black balls, then

$$P(C_{k,r,m}) = \frac{1}{N+1} \frac{k^m (N-k)^{m-r}}{N^m} = \frac{1}{N+1} \left(\frac{k}{N}\right)^r \left(1 - \frac{k}{N}\right)^{m-r}.$$

The underlying probability measure is specified by this process and hence implicitly by the priors (8.1).

To return to our question, an invocation of the theorem of total probability now yields

$$P(A_m) = \sum_{k=0}^N P(A_m | U_k) P(U_k) = \sum_{k=0}^N \frac{k^m}{(N+1)N^m}$$

for each value of m . While the probabilistic content of the problem is almost exhausted, it is useful to seek an analytical simplification of the right-hand side to make the rôles of the principal actors clearer. The sum $\sum_{k=0}^N k^m$ may be interpreted as summing the areas of $N+1$ contiguous rectangles of unit width and heights increasing from 0^m to N^m . As the function x^m increases monotonically with x , a little introspection shows that the sum may be bracketed by suitably selected areas under the curve x^m . The reader may find it useful to draw a figure. The expression on the right may hence be approximated by integrals,

$$\frac{1}{(N+1)N^m} \int_0^N x^m dx < P(A_m) < \frac{1}{(N+1)N^m} \int_1^{N+1} x^m dx,$$

which, after an evaluation of the integrals on either side and a routine algebraic clearing of terms, leads to the two-sided estimate

$$\frac{1}{m+1} \left(\frac{1}{1+N^{-1}} \right) < P(A_m) < \frac{1}{m+1} \left(\frac{(1+N^{-1})^{m+1} - N^{-m-1}}{1+N^{-1}} \right).$$

For each fixed m the expressions in the round brackets in both the upper and lower bounds tend to one as $N \rightarrow \infty$. The (unconditional) probability of observing m red balls in a row (and, possibly, more) hence takes the very informative asymptotic form

$$P(A_m) \rightarrow \frac{1}{m+1} \quad (N \rightarrow \infty). \quad (8.2)$$

From a Bayesian perspective, these probabilities may be thought to take on the rôle of priors (inherited from the uniform prior (8.1)).

Now it is clear that the occurrence of the event A_{m+1} implies the occurrence of the event A_m . Accordingly, as $N \rightarrow \infty$,

$$P(A_{m+1} | A_m) = \frac{P(A_{m+1} \cap A_m)}{P(A_m)} = \frac{P(A_{m+1})}{P(A_m)} \rightarrow \frac{m+1}{m+2} = 1 - \frac{1}{m+2},$$

and the Bayesian *a posteriori* probabilities show a markedly different behaviour: the longer the run of red balls, the more likely another red ball is to follow.

The mathematical physicist Pierre Simon de Laplace used these calculations in 1812 in his magisterial *Théorie Analytique des Probabilités* as a basis for estimating the probability that the sun will rise tomorrow. He wrote “Placing the most ancient epoch of history at five thousand years ago, or at 1,826,213 days, and the sun having risen constantly in the interval at each revolution of twenty-four hours, it is a bet of 1,826,214 to 1 that it will rise again to-morrow.” The mathematical argument for the assumed model is impeccable; the assumptions of the model themselves are open to question. As Laplace himself continues “But this number is incomparably greater for him who, recognising in the totality of phenomena the principal regulator of days and seasons, sees that nothing at the present moment can arrest the course of it.”

9 Back to the future, the Copernican principle

Suppose a phenomenon has continued unabated for m days. How much longer is it likely to last? One can imagine posing this question for a varied range of phenomena: the remaining lifetime of a wonder of the world, say the Great Wall of China, or the Great Pyramid of Cheops; the tenure of a foreign potentate; the time to extinction of a threatened species; continued investment in the exploration of outer space; the lifetime of the sun; the length of time before a champion golfer is overtaken at the top of the rankings; the persistence of a run of bad luck; the period of global warming; or the survival of *Homo sapiens*.

Conditional Probability

If we model the continued persistence of the phenomenon via Laplace's law of succession then, in view of the estimate (8.2), the probability that it persists for an additional n days tends to a well-defined limit

$$P(A_{m+n} | A_m) = \frac{P(A_{m+n})}{P(A_m)} \rightarrow \frac{m+1}{m+n+1}. \quad (9.1)$$

The reader may legitimately question the relevance of the Laplace urn model in this setting. After all, abeyant explicit information to the contrary, it appears unlikely that the lifetime of an arbitrary phenomenon is governed by repeated trials from an urn model. Key to the argument, however, is the assumption of the *uniform* Bayesian prior (8.1). Stripped of the urn metaphor, this principle can be seen in a colder, clearer light.

At the time of observation suppose that the phenomenon under investigation has continued unabated for S days and will last for an additional T days. Here S and T are to be assumed to be random quantities. Given that $S = m$, how likely is it that T exceeds n ? Clearly, one cannot make further progress without assuming something about the statistical nature of S and T . While there are many assumptions one could make on a more or less *ad hominem* basis, a profitable line of inquiry is provided by the *Copernican Principle* according to which, in the absence of evidence to the contrary, we should not think of ourselves as occupying a special position in the universe. While Copernicus fruitfully used his principle to argue against an earth-centric view of the universe, in our setting the principle may be taken to say that the point of observation of the phenomenon is uniformly distributed over its lifetime. In notation, let $R = S/(S + T)$ be the elapsed fraction of the lifetime of the phenomenon. Suppose m is any given positive integer and $0 \leq \rho \leq 1$ is a point in the unit interval. The Copernican Principle in our setting may then be construed as saying that

$$P\{R \leq \rho | S = m\} = \rho.$$

In particular, $P\{R \leq 0.25 | S = m\} = P\{R \geq 0.75 | S = m\} = 0.25$ and at the point of observation we are as likely to be in the first quartile of the lifetime of the event as in the last quartile.

The principle may be cast in terms of the future lifetime of the phenomenon. Suppose m and n are positive integers, $\rho = m/(m + n)$. Straightforward algebraic manipulation now shows that the Copernican Principle yields the conditional distribution of future life,

$$P\{T \geq n | S = m\} = \frac{m}{m+n}. \quad (9.2)$$

The range of future duration predicted by the Copernican Principle is essentially identical to that obtained from Laplace's law of succession. Indeed, if we

write $D = S + T$ for the lifetime of the phenomenon, we may restate the Copernican Principle (9.2) in the form

$$P\{D \geq m + n \mid S = m\} = \frac{m}{m + n}. \quad (9.3)$$

The reader should ponder the qualitatively similar pair (9.1,9.3) to convince herself that, but for a difference in notation, they both describe essentially the same phenomenon at work.

Suppose n_1 and n_2 are positive integers with $n_1 < n_2$. Then (9.2) says that

$$\begin{aligned} P\{n_1 \leq T < n_2 \mid S = m\} &= P\{T \geq n_1 \mid S = m\} - P\{T \geq n_2 \mid S = m\} \\ &= \frac{m}{m + n_1} - \frac{m}{m + n_2}. \end{aligned}$$

Suppose we want to make a pronouncement on the future duration of the phenomenon with a confidence in excess of $1 - \delta$ where $0 < \delta < 1$ determines the degree of our belief. Setting $n_1 = m\delta/(2 - \delta)$ and $n_2 = m(2 - \delta)/\delta$, trivial algebra now shows that

$$P\left\{m\left(\frac{\delta}{2-\delta}\right) \leq T < m\left(\frac{2-\delta}{\delta}\right) \mid S = m\right\} = \left(1 - \frac{\delta}{2}\right) - \frac{\delta}{2} = 1 - \delta$$

(ignoring irritating integer round-off factors). A 95% confidence margin is traditionally invoked as a measure of statistical significance. Accordingly, setting $\delta = 0.05$ we see that, given that a phenomenon has persisted for m days, the Copernican Principle predicts with a confidence of 95% that it will persist for at least $\frac{1}{39}m$ days and will be extinguished by at most $39m$ days. The physicist J. Richard Gott has popularised predictions based on this principle.⁶

Various diverting calculations may be based on, say, the 95% confidence bounds on future prospects arising from the Copernican Principle. For a small sample, if, following a coup, a military junta has been in power for 150 days then its days are numbered between 4 days and 16 years (a not very comforting calculation!); the Great Pyramid of Cheops is thought to have been built around 2,580 BC so, assuming a 4,600-year existence, it will continue to fascinate and intrigue for between 115 years and 180,000 years; *Homo sapiens* is thought to have evolved around 200,000 years ago in the cradle of Africa and the Copernican Principle then suggests that the species will last for another 5,128 years but will become extinct within 7.8 million years. The last example suggests a tongue-in-cheek political exhortation for space colonisation: if *Homo sapiens* is planet-bound then, by the calculation just given, it faces extinction within eight

⁶J. R. Gott, "Implications of the Copernican Principle for our future prospects", *Nature*, vol. 363, pp. 315–319, 1993.

million years or so, just like any other long-lived species, so that the best hope to stave off extinction may be to colonise space. But manned exploration of outer space has a history going back to only about 50 years which, by the Copernican Principle, suggests a future commitment (at a 95% confidence level) to exploring space lasting between one year and 1,950 years only. More to the point given that survival is at stake, the principle suggests that there is a 50% chance that humankind will cease exploring space within 150 years. Adherents of the principle will find cold comfort in the argument that our long-term survival rests upon our ability to colonise Mars, say, within the next century or so.

The Copernican Principle has achieved some notoriety in recent years, in part because of its beguiling simplicity and technicality-free predictions. But simplicity of use should not be confused with correctness. It would be foolish to use it as a predictor of the lifespan of a nonagenarian or the future tenure of a US President who has completed seven years in office—it is obvious that one cannot take the position that the observation point is not special in such cases.

The indiscriminate use of the Copernican Principle can lead to intense philosophical arguments depending on how studied one's agnosticism is. For what it is worth, common or garden variety intuition suggests that, in the absence of evidence to the contrary, a long-running phenomenon is likely to persist for a long time; new phenomena, however, are likely to be rapidly extinguished. This may explain in part why most start-ups, say, in technology, in view of late twentieth-century experience, are rapidly extinguished but those who survive and become venerable companies of long standing seem to persist indefinitely. But this is just a statement of belief in the Laplace model of succession with a (roughly) uniform prior lurking in the background. It would be well to keep the limitations of such metaphysical arguments firmly in mind.

10 Ambiguous communication

The fundamental problem in electrical communications is the presence of interference or noise in the medium (or *channel*) which results in potentially ambiguous or error-prone communications. Cell phone users have first-hand experience of these pernicious effects in the fades and abrupt drops of calls that are part and parcel of the wireless experience. But the phenomenon is not restricted to wireless settings: channel noise is a bane of any form of electrical communications.

In modern digital communications the sender transmits a sequence of symbols from a source alphabet over a channel, the transmitted sequence engendering in turn a sequence of received symbols from some (possibly different) alphabet. I will specialise to finite alphabets for definiteness (though the concepts extend without difficulty to infinite alphabets) and consider the situation when a single symbol is transmitted.

Suppose a symbol S from a finite input alphabet $\{s_1, \dots, s_m\}$ is selected and transmitted at a given transmission epoch and occasions receipt of a symbol R taking values in some output alphabet $\{r_1, \dots, r_n\}$. In typical applications the size of the output alphabet is larger than that of the input alphabet though this is not essential for the analysis to follow. The effect of channel noise is to cause uncertainty in the received symbol R . For instance, in pulse amplitude modulation, the input alphabet consists of a finite set of electrical pulse amplitudes. Noise in the intervening medium, however, can cause a random dissipation or other alteration in pulse amplitude leading to uncertainty in what was transmitted. The fundamental question that faces the recipient in this setting is, given that a received symbol R is observed, what was the corresponding transmitted symbol S ? A *decision rule* ρ is a map which associates with each output symbol r an input symbol $\rho(r)$ which constitutes the recipient's guess as to what was transmitted. The objective of the communications engineer is to implement the objectively best decision rule.

Statistical vagaries in this setting may be modelled most naturally by positing a system of conditional probabilities $\{p(r_k | s_j), 1 \leq j \leq m, 1 \leq k \leq n\}$, where $p(r_k | s_j) := P\{R = r_k | S = s_j\}$ denotes the conditional probability that symbol r_k is received when symbol s_j is transmitted. The conditional probabilities $p(r_k | s_j)$ capture symbol-by-symbol channel uncertainty and are called the channel *transition probabilities*. The statistical picture is complete if, additionally, one ascribes to the transmitted symbol S an *a priori* probability distribution $\{a(s_j), 1 \leq j \leq m\}$, where $a(s_j) := P\{S = s_j\}$ denotes the probability that symbol s_j is selected and transmitted. The underlying sample space is now naturally the space of the mn possible transmit-receive pairs $\{(s_j, r_k), 1 \leq j \leq m, 1 \leq k \leq n\}$ with the distribution of pairs inherited from the numbers $\{a(s_j)\}$ and $\{p(r_k | s_j)\}$.

Now suppose $\rho: \{r_1, \dots, r_n\} \rightarrow \{s_1, \dots, s_m\}$ is a given decision rule. A decision error is occasioned by the use of ρ if $\rho(R) \neq S$. Let E_ρ be the event that an error is made by use of ρ . It is clear then that E_ρ consists of precisely those pairs (s, r) for which $\rho(r) \neq s$. By total probability, the probability that a decision error is made under ρ is hence given by

$$P(E_\rho) = P\{\rho(R) \neq S\} = 1 - P\{\rho(R) = S\} = 1 - \sum_{k=1}^n P\{S = \rho(r_k) | R = r_k\} P\{R = r_k\}.$$

The expression on the right places the impact of the choice of decision rule on the error probability in sharp relief. By total probability, the distribution of the received symbols $\{P\{R = r_1\}, \dots, P\{R = r_n\}\}$ is completely determined by the system of channel transition probabilities $\{p(r_k | s_j), 1 \leq j \leq m, 1 \leq k \leq n\}$ and the *a priori* input symbol probabilities $\{a(s_j), 1 \leq j \leq m\}$ and is in consequence unaffected by the choice of decision rule. The choice of ρ only manifests itself in the *a posteriori* probabilities $P\{S = \rho(r_k) | R = r_k\}$ in the sum on the right. It follows that any decision rule ρ^* that maximises the symbol *a posteriori* probabilities

has the smallest error probability. Formally, for each $k = 1, \dots, n$, set $\rho^*(r_k) = s_{k^*}$ if $P\{S = s_{k^*} | R = r_k\} \geq P\{S = s_j | R = r_k\}$ for each j . (Ties may be broken in any convenient manner.) Then $P(E_{\rho^*}) \leq P(E_\rho)$ for any choice of decision rule ρ . Decision rules ρ^* that maximise the *a posteriori* probabilities are also said to be *Bayes optimal*.

Bayes's rule makes it a simple matter to explicitly specify an optimal decision rule given the channel transition probabilities and the *a priori* symbol probabilities. For each k , the *a posteriori* probabilities are given as j varies by

$$P\{S = s_j | R = r_k\} = \frac{P\{R = r_k | S = s_j\} P\{S = s_j\}}{P\{R = r_k\}} = \frac{p(r_k | s_j) a(s_j)}{\sum_{i=1}^m p(r_k | s_i) a(s_i)}.$$

As the denominator on the right does not depend on j , it follows that to maximise the *a posteriori* probability $P\{S = s_j | R = r_k\}$ for a given k it suffices to maximise the term $p(r_k | s_j) a(s_j)$ as j varies. Accordingly, for each value of k , a maximum *a posteriori* probability decision rule ρ^* implements a map $\rho^*: r_k \mapsto s_{k^*}$ where $k^* = \arg \max_{j=1, \dots, m} p(r_k | s_j) a(s_j)$.

11 Problems

1. *Dice.* If three dice are thrown what is the probability that one shows a 6 given that no two show the same face? Repeat for n dice where $2 \leq n \leq 6$.
2. *Keys.* An individual somewhat the worse for drink has a collection of n keys of which one fits his door. He tries the keys at random discarding keys as they fail. What is the probability that he succeeds on the r th trial?
3. Let π_1, \dots, π_n be a random permutation of the numbers $1, \dots, n$. If you are told that $\pi_k > \pi_1, \dots, \pi_k > \pi_{k-1}$, what is the probability that $\pi_k = n$?
4. *Balls and urns.* Four balls are placed successively in four urns, all arrangements being equally probable. Given that the first two balls are in different urns, what is the probability that one urn contains exactly three balls?
5. *Bridge.* North and South have ten trumps between them, trumps being cards of a specified suit. (a) Determine the probability that the three remaining trumps are in the same hand, that is to say, either East or West has no trumps. (b) If it is known that the king of trumps is included among the missing three, what is the probability that he is "unguarded", that is to say, one player has the king and the other the remaining two trumps?
6. *Defect origin.* Three semiconductor chip foundries, say, A, B, and C, produce 25%, 35%, and 40%, respectively, of the chips in the market. Defective chips are produced by these foundries at a rate of 5%, 4%, and 2%, respectively. A semiconductor chip is randomly selected from the market and found to be defective. What are the probabilities that it originated at A, B, and C, respectively?
7. *The ballot problem.* Suppose candidate \mathfrak{A} wins n votes and candidate \mathfrak{B} wins m votes in a two-party election. Let $P_{n,m}$ denote the probability that \mathfrak{A} leads \mathfrak{B} throughout the count. Determine the probabilities $P_{n,1}$ and $P_{n,2}$ by direct combinatorial methods.

8. Parity. A coin with success probability p is tossed repeatedly. Let P_n be the probability that there are an even number of successes in n tosses. Write down a recurrence for P_n by conditioning and hence determine P_n analytically for each n . Thence verify the identity $\sum_{k \text{ even}} \binom{n}{k} = 2^{n-1}$.

9. Uncertain searches. The Snitch in the game of Quidditch is famously elusive. It is hidden in one of n boxes, location uncertain, and is in box k with probability p_k . If it is in box k , a search of the box will only reveal it with probability α_k . Determine the conditional probability that it is in box k given that a search of box j has not revealed it. [Hint: Consider the cases $j = k$ and $j \neq k$ separately.]

10. Total probability. Prove from first principles that

$$P(A | H) = P(A | B \cap H) P(B | H) + P(A | B^c \cap H) P(B^c | H)$$

whenever $B \cap H$ and $B^c \cap H$ have non-zero probability.

11. Seven balls are distributed randomly in seven urns. If exactly two urns are empty, show that the conditional probability of a triple occupancy of some urn equals $1/4$.

12. Die A has four red and two white faces; die B has two red and four white faces. A coin is tossed once privately, out of sight of the player; if it turns up heads then die A is selected and tossed repeatedly; if it turns up tails then die B is selected and tossed repeatedly. (a) What is the probability that the n th throw of the selected die results in a red face? (b) Given that the first two throws are red, what is the probability that the third throw results in a red face? (c) Suppose that the first n throws resulted in successive red faces. What is the probability that the selected die is A?

13. A man possesses five coins, two of which are double-headed, one is double-tailed, and two are normal. He shuts his eyes, picks a coin at random, and tosses it. (a) What is the probability that the lower face of the coin is a head? (b) He opens his eyes and sees that the coin is showing heads; what is the probability that the lower face is a head? He shuts his eyes again, and tosses the coin again. (c) What is the probability that the lower face is a head? (d) He opens his eyes and sees that the coin is showing heads; what is the probability that the lower face is a head? He discards the coin, picks another at random, and tosses it. (e) What is the probability that it showed heads?

14. Random selection. In a school community families have at least one child and no more than k children. Let n_j denote the number of families having j children and let $n_1 + \dots + n_k = m$ be the number of families in the community. A child representative is picked by first selecting a family at random from the m families and then picking a child at random from the children of that family. What is the probability that the chosen child is a first born? If, alternatively, a child is chosen directly by random selection from the pool of children, what is the probability that the child is a first born? Show that the first procedure leads to a larger probability of selecting a first born. [Hint: It will be necessary to prove the identity $\sum_{i=1}^k i n_i \sum_{j=1}^k \frac{n_j}{j} \geq \sum_{i=1}^k n_i \sum_{j=1}^k n_j$ to establish the final part of the problem.]

15. Random number of dice. A random number of dice, say, N in number is selected from a large collection of dice. The number N has distribution $P\{N = k\} = 2^{-k}$ for $k \geq 1$. The selected dice are thrown and their scores added to form the sum S . Determine the following probabilities: (a) $N = 2$ given that $S = 4$; (b) $S = 4$ given that N is even; (c) the largest number shown by any die is r .

Conditional Probability

16. Suppose A , B , and C are events of strictly positive probability in some probability space. If $P(A \mid C) > P(B \mid C)$ and $P(A \mid C^c) > P(B \mid C^c)$, is it true that $P(A) > P(B)$? If $P(A \mid C) > P(A \mid C^c)$ and $P(B \mid C) > P(B \mid C^c)$, is it true that $P(A \cap B \mid C) > P(A \cap B \mid C^c)$? [Hint: Consider an experiment involving rolling a pair of dice.]
17. *Pólya's urn scheme.* In the urn model of Section 5, what is the probability that the first ball selected was black given that the second ball selected was black?
18. *Continuation.* Show by induction that the probability of selecting a black ball in the k th drawing is invariant with respect to k .
19. *Continuation.* Suppose $m < n$. Show that the probability that black balls are drawn on the m th and n th trials is $s(s+a)/(r+s)(r+s+a)$; the probability that a black ball is drawn on the m th trial and a red ball is drawn on the n th trial is $rs/(r+s)(r+s+a)$.
20. An urn contains n balls, each of a different colour. Balls are drawn randomly from the urn, with replacement. Assuming that each colour is equally likely to be selected, determine an explicit expression for the probability $P = P(M, n)$ that after $M \geq n$ successive draws one or more of the colours has yet to be seen.
21. *Continuation.* Set $M = \lfloor n \log n \rfloor$, the logarithm to base e . Find an asymptotic estimate for the probability $P(M, n)$ as $n \rightarrow \infty$. What can you conclude for large n ?
22. *The Bernoulli model of diffusion.* Another model of diffusion was suggested by D. Bernoulli. As in the Ehrenfest model, consider two chambers separated by a permeable membrane. A total of N red balls and N black balls are distributed between the two chambers in such a way that both chambers contain exactly N balls. At each succeeding epoch a ball is randomly selected from each chamber and exchanged. The state of the system at any epoch may be represented by the number of red balls in the left chamber. Determine the transition probabilities p_{jk} for this model of diffusion and thence the stationary probabilities u_k . [Hint: Use the combinatorial identity of Problem VIII.2.]
23. Three players a , b , and c take turns at a game in a sequence of rounds. In the first round a plays b while c waits for her turn. The winner of the first round plays c in the second round while the loser skips the round and waits on the sideline. This process is continued with the winner of each round going on to play the next round against the person on the sideline with the loser of the round skipping the next turn. The game terminates when a player wins two rounds in succession. Suppose that in each round each of the two participants, whoever they may be, has probability $1/2$ of winning unaffected by the results of previous rounds. Let x_n , y_n , and z_n be the conditional probabilities that the winner, loser, and bystander, respectively, in the n th round wins the game given that the game does not terminate at the n th round. (a) Show that $x_n = \frac{1}{2} + \frac{1}{2}y_{n+1}$, $y_n = \frac{1}{2}z_{n+1}$, and $z_n = \frac{1}{2}x_{n+1}$. (b) By a direct argument conclude that, in reality, $x_n = x$, $y_n = y$, and $z_n = z$ are independent of n and determine them. (c) Conclude that a wins the game with probability $5/14$.
24. *Gambler's ruin.* A gambler starts with k units of money and gambles in a sequence of trials. At each trial he either wins one unit or loses one unit of money, each of the possibilities having probability $1/2$ independent of past history. The gambler plays until his fortune reaches N at which point he leaves with his gains or until his fortune becomes 0 at which point he is ruined and leaves in tears swearing never to betray his mother's trust again. Determine the probability q_k that he is bankrupted.

25. *Families.* Let the probability p_n that a family has exactly n children be αp^n when $n \geq 1$ with $p_0 = 1 - \alpha p(1 + p + p^2 + \dots)$. Suppose that all sex distributions of the n children have the same probability. Show that for $k \geq 1$ the probability that a family has exactly k boys is $2\alpha p^k / (2 - p)^{k+1}$.

26. *Seating misadventures.* A plane with seating for n passengers is fully booked with each of n passengers having a reserved seat. The passengers file in and the first one in sits at an incorrect seat selected at random from the $n - 1$ seats assigned to the other passengers. As each of the subsequent passengers file in they sit in their allotted seat if it is available and sit in a randomly selected empty seat if their seat is taken. What is the probability f_n that the last passenger finds his allotted seat free?

27. *Laplace's law of succession.* If r red balls and $m - r$ black balls appear in m successive draws in the Laplace urn model, show that the probability that the $(m + 1)$ th draw is red tends asymptotically with the number of urns to $(r + 1)/(m + 2)$. [Hint: Prove and use the identity $I(n, m) := \int_0^1 x^n (1 - x)^m dx = n!m!/(n + m + 1)!$ by integrating by parts to obtain a recurrence and starting with the known base $I(n, 0) = 1/(n + 1)$ to set up an induction over m .]

28. *The marriage problem.* A princess is advised that she has n suitors whose desirability as consort may be rank-ordered from low to high. She does not know their rank-ordering ahead of time but can evaluate the relative desirability of each suitor by an interview. She interviews the suitors sequentially in a random order and at the conclusion of each interview either accepts the suitor as consort or casts him into the wilderness. Her decisions are final; once a suitor is rejected she cannot reclaim him, and once a suitor is accepted then the remaining suitors are dismissed. If she has run through the list and rejected the first $n - 1$ suitors she sees then she accepts the last suitor whatever his relative desirability. How should she proceed? This is the prototypical problem of *optimal stopping* in *sequential decision theory*. A predetermined stopping point m has a chance only of $1/n$ of catching the best consort. But consider the following strategy: the princess fixes a number $1 \leq r \leq n - 1$ in advance, discards the first r suitors while keeping track of the desirability of the best candidate in the discarded group, and then accepts the first suitor she encounters in the subsequent interviews whose desirability is higher than the best of the discarded set of r suitors (if there are any such else she accepts the last suitor on the list). Determine the probability $P_n(r)$ that she selects the most desirable suitor. This problem is due to Merrill M. Flood.⁷ [Hint: Condition on the location of the best suitor in the sequence.]

29. *Continuation.* By approximating a sum by integrals show that, as $n \rightarrow \infty$, she should select r as the closest integer to $(n - 1)/e$ and that with this choice her probability of accepting the best consort is approximately $1/e$, almost independent of n .

30. *Optimal decisions.* A communications channel has input symbol set $\{0, 1\}$ and output symbol set $\{a, b, c\}$. In a single use of the channel a random input symbol S governed by the *a priori* probabilities $P\{S = 0\} = 0.7$ and $P\{S = 1\} = 0.3$ is selected and transmitted.

$p(r s)$	a	b	c
0	0.7	0.2	0.1
1	0.3	0.2	0.5

⁷Variations on this theme have found applications in problems ranging from college admissions to spectrum sharing in cognitive radio. For a classical perspective see D. Gale and L. S. Shapley, "College admissions and the stability of marriage", *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

Conditional Probability

It engenders an output symbol R governed by the transition probabilities $p(r | s) := P\{R = r | S = s\}$ given in the adjacent table. Determine the optimal decision rule $\rho^*: \{a, b, c\} \rightarrow \{0, 1\}$ that minimises the probability of decision error and determine this minimum error probability.

31. Continuation, minimax decisions. There are eight possible decision rules for the communications channel of the previous problem. For the given matrix of channel conditional probabilities, allow the input symbol probability $q := P\{S = 0\}$ to vary over all choices in the unit interval and plot the probability of error for each decision rule versus q . Each of the eight decision rules ρ has a maximum probability of error which occurs for some least favourable *a priori* probability q_ρ for the input symbol 0. The decision rule which has the smallest maximum probability of error is called the *minimax decision rule*. Which of the eight rules is minimax?

III

A First Look at Independence

The concept of statistical independence is fundamental in the theory of probability. The distinctive and rich intuitive content of the theory and its link to observations in physical experiments is provided by the connection between the narrow, but well-defined and formal, notion of statistical independence as a rule of multiplication of probabilities, and the vague, but broad, sense of physical “independence” connoting unrelated events. In view of its great importance it will be well worth our while to spend time understanding the historical background, motivation, and ultimately formal definition and properties of independence. We will repeatedly return to this most basic and central issue from many different vantage points throughout the book.

C 1, 4
A 2, 3
F 5

1 A rule of products

Suppose A is a set of m elements and B a set of n elements. How many ways can we form ordered pairs of elements (a, b) with $a \in A$ and $b \in B$? If a and b may be specified independently (in our common or garden understanding of the word), then for every choice of a we may specify m choices for b whence there are a total of

$$\underbrace{m + m + \cdots + m}_{n \text{ terms}} = mn$$

distinct choices for the pairs (a, b) . Elementary, but this underscores a critical arithmetic relationship: *possibilities multiply under independent selection*.

EXAMPLES: 1) *Coins*. If a fair coin is tossed twice there are two possibilities, \mathfrak{H} and \mathfrak{T} , for each of the tosses and $4 = 2 \times 2$ possible outcomes, $\mathfrak{H}\mathfrak{H}$, $\mathfrak{H}\mathfrak{T}$, $\mathfrak{T}\mathfrak{H}$, and $\mathfrak{T}\mathfrak{T}$, for the experiment. The probability of a head on the first trial is $1/2$, as is the probability of a tail in the second trial. The probability that we observe the outcome \mathfrak{HT} is $\frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$.

2) *Dice*. If six dice are tossed there are 6^6 possible outcomes. Exactly three out of the six possible outcomes of any given toss yield an even face so that the

probability of an even face is $3/6 = 1/2$. The event that all six tosses result in an even face has 3^6 favourable outcomes and hence probability $\frac{3^6}{6^6} = \left(\frac{1}{2}\right)^6$.

3) *Urns.* An urn has 1 black ball, 1 green ball, 2 blue balls, and 4 red balls, all balls considered distinguishable. If one ball is removed at random from the urn, the probability of drawing a black, green, blue, or red ball is $1/8$, $1/8$, $1/4$, and $1/2$, respectively. If four balls are drawn with replacement, the total number of possible outcomes is 8^4 ; the number of outcomes favourable to the event that the draws are, in order, black, green, blue, and red is $1 \cdot 1 \cdot 2 \cdot 4 = 8$. The probability of this event is $\frac{8}{8^4} = \frac{1}{8} \cdot \frac{1}{8} \cdot \frac{1}{4} \cdot \frac{1}{2}$. ▶

The idea of statistical independence is a codification of, and can eventually be seen to be a very substantial generalisation of, this very simple and intuitive idea of independent selection. *At its most basic level statistical (or stochastic) independence may be defined as a rule of multiplication of probabilities.* The reader will note that we have already tacitly used this concept of independence in building up the probability model in Laplace's law of succession in Section II.8.

Let us start by defining formally what it means for a collection of events to be statistically independent. From now on we drop the qualifiers "statistical" or "stochastic" when we talk about independence in a probabilistic setting. In the sequel, all events are subsets of some underlying, abstract probability space. We will start with the simplest setting of a pair of events for which the intuition seems reasonably secure and build up from there to more and more complex settings.

DEFINITION 1 *Two events A and B are independent if $\mathbf{P}(A \cap B) = \mathbf{P}(A) \mathbf{P}(B)$.*

In particular, if $\mathbf{P}(B) \neq 0$, then the independence of A and B is equivalent to the intuitive statement $\mathbf{P}(A | B) = \mathbf{P}(A)$, i.e., A "does not depend on" B. Likewise, independence implies $\mathbf{P}(B | A) = \mathbf{P}(B)$ (again provided A has non-zero probability, else of course the conditional probability is undefined).

If A and B are independent then, informally speaking, B contains no information about A. In which case B^c should also contain no information about A, else the probability model would be absurd. But this is easily verified via additivity as

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

Thus, if A and B are independent, then so are A and B^c . Of course, it also follows that A^c and B are independent, as are A^c and B^c .

The basic definition adapts easily to a formal idea of independence in conditional settings: two events A and B are *conditionally independent given an event C* if $\mathbf{P}(A \cap B | C) = \mathbf{P}(A | C) \mathbf{P}(B | C)$. No new ideas are needed here as we recall that conditioning on C merely induces the new probability measure $\mathbf{P}(\cdot | C)$; conditional independence is merely a product rule with respect to

this measure. Care should be taken not to read too much into the definition: conditional independence does not, in general, imply independence, or vice versa.

We shall accept without further comment that subsequent definitions can be adapted to conditional settings in a similar fashion.

Additional conditions are needed when we move from pairs to triples of events.

DEFINITION 2 *Three events A, B, and C are independent if all four of the following conditions are satisfied:*

$$\begin{aligned} \mathbf{P}(A \cap B) &= \mathbf{P}(A) \mathbf{P}(B), \\ \mathbf{P}(A \cap C) &= \mathbf{P}(A) \mathbf{P}(C), \\ \mathbf{P}(B \cap C) &= \mathbf{P}(B) \mathbf{P}(C), \\ \mathbf{P}(A \cap B \cap C) &= \mathbf{P}(A) \mathbf{P}(B) \mathbf{P}(C). \end{aligned} \tag{1.1}$$

In other words, A, B, and C have to be mutually *pairwise independent* (that is to say, the first three conditions of (1.1) have to hold) *and*, in addition, the fourth condition dealing with the conjunction of all three events has also to be satisfied. While it may appear that it is unnecessary to specify all four conditions, as we shall see shortly, it is an unfortunate fact but true that no three of these conditions imply the fourth.

Of course, if A, B, and C are independent, one anticipates, for instance, that $A \cup B$ is also independent of C. I will leave the demonstration as an exercise.

The conditions to be checked multiply very quickly when the number of events increases.

DEFINITION 3 *Events A_1, \dots, A_n are independent if*

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \mathbf{P}(A_{i_1}) \mathbf{P}(A_{i_2}) \dots \mathbf{P}(A_{i_k})$$

for every integer k = 1, ..., n, and every choice of indices $1 \leq i_1 < i_2 < \dots < i_k \leq n$.

When $k = 1$ the condition is trivial. This still leaves $\binom{n}{2}$ conditions to be checked for pairs, $\binom{n}{3}$ for triples, and so on, resulting in a total of $\binom{n}{2} + \binom{n}{3} + \dots + \binom{n}{n} = 2^n - n - 1$ non-trivial conditions. Independence places very strenuous requirements on the interrelationships between events.

And, finally, we have the general

DEFINITION 4 *A countable collection of events $\{A_i, i \geq 1\}$ is independent if*

$$\mathbf{P}(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}) = \mathbf{P}(A_{i_1}) \mathbf{P}(A_{i_2}) \dots \mathbf{P}(A_{i_k})$$

for every k and every distinct choice of k positive integers i_1, \dots, i_k .

In other words, a sequence of events $\{A_i, i \geq 1\}$ is independent if the product property is satisfied for every *finite* subcollection of events.

2 What price intuition?

At the simplest levels, the definition of independence as a rule of products appears natural and unexceptionable.

EXAMPLES: 1) *Cards*. A card is drawn randomly from a 52-card pack. The event that it is an ace has four favourable outcomes, hence probability $4/52 = 1/13$. The event that the suit of the card is spades has 13 favourable outcomes and probability $13/52 = 1/4$. We may anticipate by symmetry that these two events are independent and, indeed, the conjoined event that the card is the ace of spades has a single favourable outcome and probability $\frac{1}{52} = \frac{1}{13} \cdot \frac{1}{4}$.

2) *Relative ranks*. Suppose all permutations of rankings of four students a , b , c , and d are equally likely. The events A that a ranks ahead of d and B that b ranks ahead of c are intuitively independent. Indeed, as is easy to verify, $\mathbf{P}(A) = \mathbf{P}(B) = 1/2$ and $\mathbf{P}(A \cap B) = 1/4$.

3) *Dice*. Two dice are thrown. The event A that the first die shows an ace has probability $1/6$ while the event B that the sum of the face values of the two dice is odd has probability $1/2$ (as there are 18 odd-even or even-odd combinations). And as $A \cap B = \{(1, 2), (1, 4), (1, 6)\}$ has probability $\frac{1}{12} = \frac{1}{6} \cdot \frac{1}{2}$, the events are independent. ►

It is quite satisfactory in the simple situations considered above that native intuition serves in lieu of a formal calculation. This suggests that the mathematical definition of independence captures the “right” abstraction. Intuition, however, is not a reliable guide and one needs to turn to actual computation to verify whether systems of events (or random variables) are indeed, formally, independent. Cautionary examples are provided below.

EXAMPLES: 4) *Male and female progeny in families*. Suppose a family has $n \geq 2$ children. All arrangements of the sexes of the children are assumed to be equally likely. Let A be the event that the family has at most one girl. Let B be the event that the family has children of both sexes. Are A and B independent?

By direct enumeration, $\mathbf{P}(A) = \frac{1}{2^n} + \frac{n}{2^n}$ and $\mathbf{P}(B) = 1 - \frac{2}{2^n} = 1 - \frac{1}{2^{n-1}}$. On the other hand, $A \cap B$ is the event that the family has exactly one girl so that $\mathbf{P}(A \cap B) = \frac{n}{2^n}$. Thus, A and B are independent if, and only if,

$$\frac{(n+1)}{2^n} \left(1 - \frac{1}{2^{n-1}}\right) = \frac{n}{2^n}, \quad \text{or, in other words, } n+1 = 2^{n-1}.$$

But this equation can be satisfied if, and only if, $n = 3$. Consequently, A and B are independent if $n = 3$ but dependent if $n \neq 3$.

5) *Conditional independence*. We are provided with two coins, one of which has probability $1/6$ for heads, while the other has probability $5/6$ for heads. One

of the two is picked at random and tossed twice. Let A_1 and A_2 be the events that the first toss and the second toss, respectively, are heads and let C_1 denote the event that the first coin is picked. The problem statement makes clear that, conditioned on the occurrence of the event C_1 , we have $P(A_1 | C_1) = P(A_2 | C_1) = 1/6$ and $P(A_1 \cap A_2 | C_1) = 1/36$, that is to say, the events A_1 and A_2 are conditionally independent given that C_1 has occurred.

On the other hand, by conditioning on which coin is picked, by the theorem of total probability it is apparent that $P(A_1) = P(A_2) = \frac{1}{2} \cdot \frac{1}{6} + \frac{1}{2} \cdot \frac{5}{6} = \frac{1}{2}$ while $P(A_1 \cap A_2) = \frac{1}{2} \cdot \left(\frac{1}{6}\right)^2 + \frac{1}{2} \cdot \left(\frac{5}{6}\right)^2 = \frac{13}{36} > \frac{1}{4}$ and it is now clear that A_1 and A_2 are not (unconditionally) independent.

Random processes taking divergent paths based on random initial conditions, of which this is a very elementary illustration, have applications in a variety of settings in genetics, finance, and stochastic theory. The reader will recall having seen an earlier example in Laplace's law of succession. ►

The fact that there are unexpected challenges to intuition even in the case of two events is a bit of a blow and suggests that one should approach independence in more general situations with some care. Let us consider the case of three events.

At first blush, the rule of products specified in, say, (1.1) seems excessive. Surely, the last condition, apparently the most restrictive, should imply the first three; or, alternatively, the first three conditions should imply the fourth. Not so, as S. N. Bernstein demonstrated. Begin with an example where all four conditions are satisfied.

EXAMPLES: 6) *Independence.* Consider the results of three consecutive tosses of a fair coin. Let A be the event that the first toss results in a head, B the event that the second toss results in a head, and C the event that the third toss results in a head. Writing the outcomes of the three tosses in succession we may identify the sample space as consisting of the eight points

$$\Omega = \{\text{TTT}, \text{TTH}, \text{THH}, \text{HTT}, \text{HTH}, \text{HHT}, \text{HHT}, \text{HHH}\}.$$

The events A , B , and C are, accordingly, $A = \{\text{HTT}, \text{HTH}, \text{HHT}, \text{HHH}\}$, $B = \{\text{THH}, \text{THH}, \text{HHT}, \text{HHH}\}$, and $C = \{\text{TTT}, \text{TTH}, \text{HTH}, \text{HHT}\}$. It is natural and intuitive to assign probability $1/8$ to each sample point. (The tosses are "independent" and the coin is "fair".) As A , B , and C each contain four sample points, each has probability $1/2$. Any two of these share exactly two sample points whence all pairwise intersections have probability $1/4$. Finally the intersection of all three of these events contains the single point HHH , hence has probability $1/8$. *Ergo:* The events A , B , and C are independent.

The Venn diagram of Figure 1(a) schematically shows the events A , B , and C . We can construct another abstract probability experiment reflecting this

A First Look at Independence

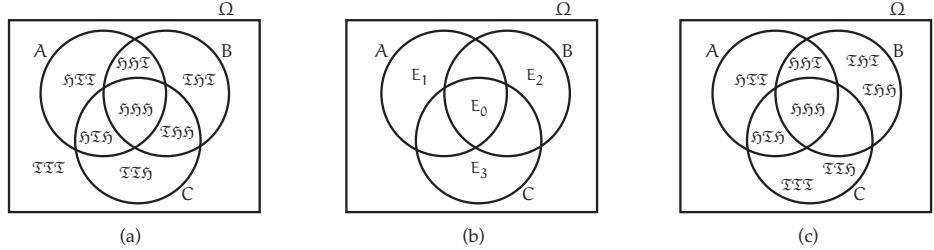


Figure 1: Venn diagrams for three experiments.

example quite simply by picking a sample space with eight points and placing one sample point in each of the eight disjoint regions in the Venn diagram. Assigning each of the sample points equal probability $1/8$ completes the job. Variations on the theme, however, break the symmetric structure fatally.

7) *Dependency I.* Here's the classical example of dependency due to S. N. Bernstein. Toss a pair of dice distinguishable, say, by colour, Red and Green. Let A be the event that the Red die shows an odd face, B the event that the Green die shows an odd face, and C the event that the sum of the two face values of the dice is an odd number. Then A consists of the 18 sample points of the form $(i, *)$ where i is 1, 3, or 5 and $*$ connotes any of the six possible outcomes for the Green die. Thus, A has probability $18/36 = 1/2$. Likewise, B has probability $1/2$. Finally, C contains exactly those sample points (i, j) where either i is odd and j is even, or i is even and j is odd. Thus, C also contains exactly 18 sample points whence it also has probability $18/36 = 1/2$.

Now $A \cap B$ consists of the sample points (i, j) where both i and j are odd. It follows that $A \cap B$ contains exactly nine sample points whence $P(A \cap B) = 9/36 = 1/4$. Similarly, $A \cap C$ consists of the sample points (i, j) where i is odd and j is even. Thus, $A \cap C$ contains nine sample points, hence has probability also $1/4$. Finally, $B \cap C$ consists of the nine sample points (i, j) where i is even and j is odd whence $B \cap C$ has probability also $1/4$. It follows that *the events A, B, and C are pairwise independent*.

On the other hand, the intersection of all three of the events A , B , and C is empty. Thus, $P(A \cap B \cap C) = 0$. *The events A, B, and C are not independent even though they are pairwise independent.*

In a variation on this theme, consider the sample space Ω consisting of the four points E_0 , E_1 , E_2 , and E_3 , with each of the four sample points carrying equal probability $1/4$. Set $A = \{E_0, E_1\}$, $B = \{E_0, E_2\}$, and $C = \{E_0, E_3\}$. This results in the situation shown in Figure 1(b). The events A , B , and C each have probability $1/2$ and are *pairwise independent*, $P(A \cap B) = P(A \cap C) = P(B \cap C) = 1/4$. However, $P(A \cap B \cap C) = 1/4 \neq 1/8 = P(A)P(B)P(C)$. Thus, A , B , and C are *not independent*.

8) *Dependency II.* Here's another experiment along the same vein. Consider the sample space comprised of the nine triples of the letters a , b , and c given by

$$\Omega = \{abc, acb, bac, bca, cab, cba, aaa, bbb, ccc\},$$

each triple accorded probability $1/9$. Let A , B , and C be the events that the letter a occurs as the first element, b occurs as the second element, and c occurs as the third element, respectively, of a randomly drawn triple. As there are precisely three triples starting with a , it is clear that $P(A) = 1/3$, and likewise, $P(B) = 1/3$ and $P(C) = 1/3$. The events A and B can occur together only if the triple abc is drawn whence $P(A \cap B) = 1/9$. By the same reasoning, $P(A \cap C) = 1/9$ and $P(B \cap C) = 1/9$. It follows that the events A , B , and C are pairwise independent. However, the events A , B , and C can occur conjointly also only if the triple abc is drawn. It follows that $P(A \cap B \cap C) = 1/9 \neq 1/27$ and again, *the events A, B, and C are not (jointly) independent even though they are pairwise independent.*

9) *Dependency III.* It is perhaps even more confounding that the fourth condition of (1.1) also does not in itself imply the others. Consider anew three tosses of a fair coin and the events

$$\begin{aligned} A &= \{\text{HHT}, \text{HHT}, \text{HHH}, \text{HTH}\}, \\ B &= \{\text{HHT}, \text{HHT}, \text{HTT}, \text{THH}\}, \\ C &= \{\text{HTH}, \text{HHT}, \text{TTT}, \text{TTT}\}. \end{aligned}$$

It is clear that A , B , and C each have probability $1/2$. As the three events only have the outcome HHH in common, it follows that $P(A \cap B \cap C) = P\{\text{HHH}\} = 1/8$. However, the pairwise product rule breaks down. While $P(A \cap B) = P\{\text{HHT}, \text{HHT}\} = 1/4 = P(A)P(B)$, and $P(A \cap C) = P\{\text{HTH}, \text{HHT}\} = 1/4 = P(A)P(C)$, we come a cropper with $P(B \cap C) = P\{\text{HTT}\} = 1/8 \neq P(B)P(C)$ and the events are not (jointly) independent. The situation is illustrated in the Venn diagram of Figure 1(c). ▶

In applications one frequently has the feeling that certain events are independent. Such intuition should not be gainsaid—it frequently provides the jump-off point to an attack on the problem. The examples given above illustrate, however, that caution is necessary; intuition should be bolstered by explicit verification of the product rules whenever possible.

3 An application in genetics, Hardy's law

Mendel's theory of heredity has seen its full flowering with the discovery of how hereditary information is carried by genes. The passing of genetic characteristics to offspring is a chance-driven process and provides fertile ground for

illustration of probabilistic methods. Our discussion is necessarily simplistic—the specialised nature of the subject inhibits a detailed study here, no less the author’s own very limited knowledge of the subject—but the ample scope for probabilistic models should be apparent.

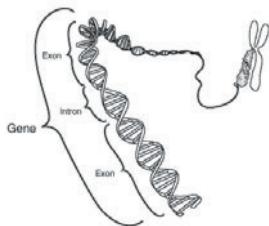


Figure 2: Genes on a chromosome.

The blueprint of hereditary features characteristic of an organism is carried within its genome, an identical copy of which is maintained within each cell. The genome consists of a complete set of genes, each of which is a specific grouping of nucleic acids, with the genes organised structurally into chromosomes. Heritable characteristics are governed by one or more genes which are located at specific points (or loci) on a chromosome. As shown in Figure 2, it may be helpful to form an image of a chromosome as a string on which are clustered a large number of beads (the genes) in a specific order.

Individual genes themselves may exist in two or more forms called alleles which are small variations in the basic molecular structure of the gene. The salient point for our purposes is that in diploid organisms (which include most higher life forms) each gene exists in pairs (the genotype) at the chromosome locus. The specific genotype of the organism ultimately informs observed variations in physical traits such as albinism, eye colour, or haemophilia that are controlled by the gene. Genotypes are classified as homozygous (meaning that the alleles forming the gene pair are identical) or heterozygous (meaning that the alleles forming the gene pair are different). We consider the simplest setting where a gene can exist as one of two alleles A and a . The organism can then belong to one of the three genotypes AA , Aa , and aa of which the first and third are homozygous and the second is heterozygous (the order is irrelevant in the arrangement Aa).

Reproductive cells (or gametes) are special in that they contain only one of the genes in the genotype. Thus, the gametes of a type AA or type aa organism contain genes of only one type, A or a , respectively, but the gametes of a type Aa organism contain A -gametes and a -gametes in equal numbers.

Genotypes of offspring are governed by chance processes. During sexual reproduction, the male gamete and the female gamete fuse to form a zygote which inherits half of its genetic material from the male and half from the female. There are then nine possible genotype combinations for the parents and three possible genotypes for the offspring that can result from gene combination. The possibilities are shown in the bipartite graph in Figure 3.

To simplify the situation, suppose that the three genotypes AA , Aa , and aa occur in the parental population among males and females in the same frequencies $u : 2v : w$ where $u + 2v + w = 1$. (The factor 2 for the heterozygous genotype Aa is purely for later algebraic convenience.) In *random mating*,

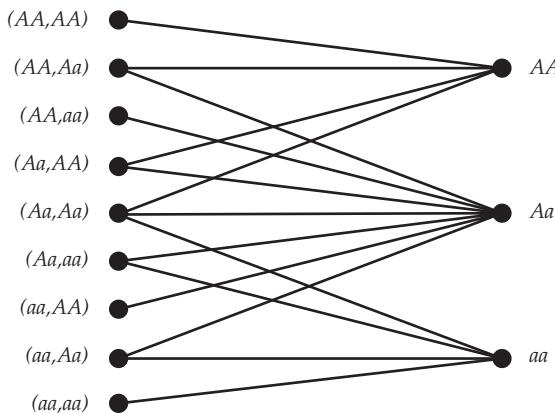


Figure 3: Genotype combinations.

which is the only case we consider here, each offspring in the filial generation is assumed to arise out of a mating of two randomly selected parents from the population (the sampling is with replacement). During mating each parent contributes a randomly selected gene chosen with equal probability $1/2$ from the gene pair constituting the parental genotype to the gene pair of the filial genotype.

Parental selection via random mating results in *a priori* genotype mixing probabilities which are simply the product of the individual genotype frequencies in the parental generation as listed in Table 1. Random gene selection

Parental pairing	<i>A priori</i> probability
(AA,AA)	u^2
(AA,Aa)	$2uv$
(AA,aa)	uw
(Aa,AA)	$2uv$
(Aa,Aa)	$4v^2$
(Aa,aa)	$2vw$
(aa,AA)	uw
(aa,Aa)	$2vw$
(aa,aa)	w^2

Table 1: Genotype combinations.

from each of the parental genotypes results then in transition probabilities to the offspring's genotype as shown in Table 2. To compute genotype frequencies in the filial generation is now a simple matter of conditioning. Write $p = u + v$

Transition probabilities	AA	Aa	aa
(AA,AA)	1	0	0
(AA,Aa)	$\frac{1}{2}$	$\frac{1}{2}$	0
(AA,aa)	0	1	0
(Aa,AA)	$\frac{1}{2}$	$\frac{1}{2}$	0
(Aa,Aa)	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
(Aa,aa)	0	$\frac{1}{2}$	$\frac{1}{2}$
(aa,AA)	0	1	0
(aa,Aa)	0	$\frac{1}{2}$	$\frac{1}{2}$
(aa,aa)	0	0	1

Table 2: Conditional genotype probabilities in the filial generation.

and $q = v + w$. We interpret p and q as the frequencies of allele A and a , respectively, in gene combination from the parental generation. An inspection of the transitions in Figure 3 shows now that an offspring belongs to the genotypes AA , Aa , and aa with probabilities u' , $2v'$, and w' , respectively, given by

$$\begin{aligned} u' &= u^2 + uv + uv + v^2 = (u + v)^2 = p^2, \\ 2v' &= uv + uw + uv + 2v^2 + vw + uw + vw = 2(uv + uw + v^2 + vw) = 2pq, \\ w' &= v^2 + vw + vw + w^2 = (v + w)^2 = q^2. \end{aligned}$$

It follows that genotype frequencies in the offspring are determined completely by the allele frequencies p and q in the parental genotypes. In particular, all choices of initial genotype frequencies u , $2v$, and w for which $u + v = p$ and $v + w = q$ will lead to the probabilities p^2 , $2pq$, and q^2 for the genotypes AA , Aa , and aa in the filial generation. What about the allele frequencies in the filial generation? As $u' + v' = p^2 + pq = p(p + q) = p$ and $v' + w' = pq + q^2 = (p + q)q = q$, it follows that the parental allele frequencies are preserved in the filial generation, and hence thereafter. In particular, if the genotypes AA , Aa , and aa exist in the parental generation in the frequencies $u : 2v : w \equiv p^2 : 2pq : q^2$ then these frequencies are transmitted unchanged as genotype probabilities in the filial generation. Genotype distributions with this property are hence called *stationary*.

In large populations, the observed genotype frequencies will be “close” to the probabilities (in a sense made precise via the law of large numbers). We’ve hence retraced the path followed by the English mathematician G. H. Hardy in 1908.¹

¹G. H. Hardy, “Mendelian proportions in a mixed population”, *Science*, vol. 28, pp. 49–50, 1908.

HARDY'S LAW *Random mating within an arbitrary parent population will produce within one generation an approximately stationary genotype distribution with gene frequencies unchanged from that of the parent population.*

Hardy's law should not be read to imply an immutability in gene frequencies. The key word is "approximate". While there is no systematic drift in any direction after the first generation, neither is there any restoring force which seeks to hold parental gene frequencies fixed. In spite of the stabilising influence of Hardy's law from generation to generation, chance fluctuations will ultimately doom the population to one of boring homogeneity if there are no changes imposed from without on the system. Fortunately, mutations, the introduction of new genetic material from outside, and many other effects mitigate against the adoption of a uniform party line (i.e., absorption into a homozygous population) in practice. *Vive la difference!*

4 Independent trials

The independence of a sequence of events entails the checking of a formidable number of product conditions. In the most important situation in practice, however, the validity of the conditions will be almost obvious and will entail no necessity for checks. This is the situation corresponding to *independent trials*.

The setting is already familiar to us in examples such as repeated coin tosses or throws of a die. We will only be concerned here with situations where individual experiments (or trials) result in a discrete set of outcomes.

EXAMPLE 1) Craps. In the popular dice game of craps a player rolls two dice and sums the face values obtained. She wins immediately if she rolls 7 or 11; if, on the other hand, she rolls 2, 3, or 12 then she loses immediately. If she rolls any of the six remaining numbers she then proceeds to roll the dice repeatedly, the game terminating at the first roll in which she either repeats her initial roll or rolls a 7; she wins if the game terminates with her initial roll, loses if it terminates with a 7. What constitutes fair odds in this setting?

It is natural to allocate equal probability $1/36$ to each pair of face values $(1,1), (1,2), \dots, (6,6)$ for the two dice. Let E_i denote the event that any given roll results in the value i for the sum of face values and let $p_i = P(E_i)$. By summing over the probabilities of the outcomes engendering each of the E_i we obtain the probabilities listed in Table I.2.1, redisplayed here for convenience:

i	2	3	4	5	6	7	8	9	10	11	12
p_i	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Let W denote the event that the player eventually wins the game. By conditioning on the outcome of the first roll, we have $P(W) = \sum_{i=2}^{12} P(W | E_i) P(E_i)$

by total probability. Now $P(W | E_2) = P(W | E_3) = P(W | E_{12}) = 0$ and $P(W | E_7) = P(W | E_{11}) = 1$. It only remains to evaluate $P(W | E_i)$ for $i \in \{4, 5, 6, 8, 9, 10\}$.

Suppose now that the game proceeds past the first roll. Let W_n denote the event that the game terminates in the n th roll. By total probability, for each $i \in \{4, 5, 6, 8, 9, 10\}$, we then have $P(W | E_i) = \sum_{n=2}^{\infty} P(W_n | E_i)$. Given that the player rolls i initially, she wins on the n th roll if, and only if, rolls 2 through $n - 1$ are neither i nor 7 and roll n is i . By independence of the rolls it follows that $P(W_n | E_i) = (1 - p_i - p_7)^{n-2} p_i$ and so, by summing a geometric series, we obtain

$$P(W | E_i) = \sum_{n=2}^{\infty} (1 - p_i - p_7)^{n-2} p_i = \frac{p_i}{p_i + p_7}.$$

Combining terms, it follows that

$$P(W) = \frac{p_4^2}{p_4 + p_7} + \frac{p_5^2}{p_5 + p_7} + \frac{p_6^2}{p_6 + p_7} + p_7 + \frac{p_8^2}{p_8 + p_7} + \frac{p_9^2}{p_9 + p_7} + \frac{p_{10}^2}{p_{10} + p_7} + p_{11},$$

and a numerical evaluation shows that the probability of winning at craps is approximately 0.4929. The somewhat artificial structure of the game is explained by the fact that the odds of winning, while approximately even are, in actuality, strictly less than 1/2: a casual player basing her observations on limited empirical evidence may well believe that the odds are fair—which explains the game's popularity—but in a long string of craps games our novice gambler will eventually lose her shirt. ▶

Suppose \mathfrak{A} and \mathfrak{B} are discrete sample spaces comprised of elements $\alpha_1, \alpha_2, \dots$, and β_1, β_2, \dots , respectively, and equipped with probabilities p_1, p_2, \dots , and q_1, q_2, \dots , respectively. The *product space* $\Omega = \mathfrak{A} \times \mathfrak{B}$ is engendered by the successive performance of the experiments corresponding to \mathfrak{A} and \mathfrak{B} . The elements of the product space are hence ordered pairs of the form (α_i, β_j) with $\alpha_i \in \mathfrak{A}$ and $\beta_j \in \mathfrak{B}$. We say that the experiments corresponding to \mathfrak{A} and \mathfrak{B} are *independent* if the probability measure associated with the product space satisfies the product rule $P\{(\alpha_i, \beta_j)\} = p_i q_j$. The fact that this does indeed yield a bona fide probability measure (called, naturally enough, a *product measure*) is verified by a trite calculation:

$$\sum_i \sum_j P\{(\alpha_i, \beta_j)\} = \sum_i \sum_j p_i q_j = \left(\sum_i p_i \right) \left(\sum_j q_j \right) = 1.$$

The case of *repeated independent trials* corresponds to the situation where \mathfrak{A} and \mathfrak{B} have the same elements and the same associated probabilities. This is the situation we are familiar with in the repeated toss of a coin or throw of a die.

Now consider independent experiments corresponding to the sample spaces \mathfrak{A} and \mathfrak{B} . Suppose A is an event in the product space which occurs if

the first element of a pair (α_i, β_j) is any of a given set of elements $\alpha_{i_1}, \alpha_{i_2}, \dots$ in \mathfrak{A} . In other words, A depends only on the result of the first experiment. It follows quickly that $P(A) = \sum_{i_k, j} P\{(\alpha_{i_k}, \beta_j)\} = \sum_{i_k, j} p_{i_k} q_j = \sum_k p_{i_k}$, as the q_j s add to one. Likewise, if B is an event in the product space which occurs if the second element of a pair (α_i, β_j) is any of a given set of elements $\beta_{j_1}, \beta_{j_2}, \dots$ in \mathfrak{B} , then B depends only on the result of the second experiment and $P(B) = \sum_l q_{j_l}$. The intersection $A \cap B$ is the event consisting of the sample points $(\alpha_{i_k}, \beta_{j_l})$ for $k, l \geq 1$. Accordingly,

$$P(A \cap B) = \sum_{i_k} \sum_{j_l} P\{(\alpha_{i_k}, \beta_{j_l})\} = \left(\sum_k p_{i_k} \right) \left(\sum_l q_{j_l} \right) = P(A) P(B),$$

and the events A and B are independent in accordance with intuition.

These considerations carry over to a sequence of independent experiments whose outcomes take values in, say, $\mathfrak{A}_1, \mathfrak{A}_2, \dots, \mathfrak{A}_n$, with \mathfrak{A}_k comprised of elements $\alpha_{k1}, \alpha_{k2}, \dots$ and equipped with probabilities p_{k1}, p_{k2}, \dots . The product space $\Omega = \mathfrak{A}_1 \times \mathfrak{A}_2 \times \dots \times \mathfrak{A}_n$ is now comprised of ordered n -tuples $(\alpha_{1i_1}, \alpha_{2i_2}, \dots, \alpha_{ni_n})$ with $\alpha_{ki_k} \in \mathfrak{A}_k$, and the product measure is given by $P\{(\alpha_{1i_1}, \alpha_{2i_2}, \dots, \alpha_{ni_n})\} = p_{1i_1} p_{2i_2} \dots p_{ni_n}$. The independence argument for pairs now generalises in a natural fashion: *if A_1, A_2, \dots, A_n are events with A_k determined exclusively by the k th trial then the events are independent*. The demonstration follows the same lines as for the case of two successive experiments.

The case of independent trials is the most developed part of the theory of probability. The simplicity and ease accorded by independence makes it worthwhile to cast experiments in this framework when possible.

EXAMPLES: 2) *Coin tosses.* The result of a coin toss is an element of the sample space $\mathfrak{A} = \{\mathfrak{H}, \mathfrak{T}\}$. Tossing a coin n times engenders a point in the n -fold product space $\Omega = \mathfrak{A} \times \dots \times \mathfrak{A}$ whose elements are the 2^n possible sequences of n heads and tails. This is the archetypal product space of repeated independent trials. It repays careful study as it captures much of the structure of general product spaces and we will be returning to it frequently.

3) *Permutations.* Consider the sample space comprised of the $n!$ permutations of elements a_1, a_2, \dots, a_n , each permutation being equally likely. Any given permutation $(a_{i_1}, a_{i_2}, \dots, a_{i_n})$ can be specified by the following $(n - 1)$ -step sequential procedure.

Begin by writing down a_1 . Then write a_2 to the left or right of a_1 depending on its location in the permutation. As one-half of all permutations have a_2 to the left of a_1 and one-half have a_2 to the right of a_1 , this corresponds to an experiment \mathfrak{A}_1 with two equally likely choices. Our next choice deals with the positioning of a_3 . There are three possibilities: a_3 is either to the left of both a_1 and a_2 , or a_3 lies between a_1 and a_2 , or a_3 lies to the right of both a_1 and a_2 . Again permutations involving these three configurations are all equally likely and the placing of a_3 vis à vis a_1 and a_2 corresponds to an experiment

\mathfrak{A}_2 with three equally likely alternatives. Proceeding in this fashion, suppose the elements a_1 through a_{k-1} have been placed. They will exist in some order $(b_1, b_2, \dots, b_{k-1})$ determined by their relative positions in the original permutation. There are now k possible locations for a_k , one on either end buttressing $k - 2$ interstices between the a_i , and the relative orientation of a_k vis à vis a_1 through a_{k-1} determines where it should be placed. The k possible locations for a_k determine a discrete sample space \mathfrak{A}_{k-1} and as all configurations for a_k with respect to a_1 through a_{k-1} are equally likely, the elements of \mathfrak{A}_{k-1} have equal probability $1/k$. The process culminates with the placement of element a_n in one of the n locations available after placing elements a_1 through a_{n-1} . The procedure just described shows that *we are dealing with the product space $\mathfrak{A}_1 \times \mathfrak{A}_2 \times \dots \times \mathfrak{A}_{n-1}$ over $n - 1$ independent trials*. Each permutation of the elements is in one-to-one correspondence with an outcome arising out of these $n - 1$ independent trials, hence has probability $\frac{1}{2} \cdot \frac{1}{3} \cdots \frac{1}{n} = \frac{1}{n!}$ as required.

4) *Sampling without replacement.* The process of specifying the ordered sample (a_2, a_5, a_3) from the population $\{a_1, a_2, a_3, a_4, a_5\}$ may be spelled out sequentially as follows. First select a_2 from the original population of five elements $\{a_1, a_2, a_3, a_4, a_5\}$. Remove a_2 from the original population and select a_5 from $\{a_1, a_3, a_4, a_5\}$. Then remove a_5 from the population and select a_3 from $\{a_1, a_3, a_4\}$. This is equivalent to the following procedure. Order the elements of the population from 1 through 5. Specify the second element, remove it, and renumber the remaining elements from 1 through 4 preserving the original order. Specify the fourth element of this new population, remove it, and renumber the remaining elements from 1 through 3, again preserving the order. Finally, specify the second element of the remaining population. With this understanding, it is clear that the sequence $(2, 4, 2)$ is entirely equivalent to the ordered sample (a_2, a_5, a_3) .

The procedure may of course be extended to a sample and population of any size: specifying an ordered sample of size r from a population of size n is equivalent to sequentially specifying an index from 1 through n , then an index from 1 through $n - 1$, and so on, culminating with an index from 1 through $n - r + 1$. Thus, any ordered r -sample $(a_{i_1}, a_{i_2}, \dots, a_{i_r})$ drawn from a population of size n is equivalent to a sequence of integers (j_1, j_2, \dots, j_r) with each j_k in the range $1 \leq j_k \leq n - k + 1$. Equivalently, if we write $\mathfrak{B}_k = \{1, 2, \dots, n - k + 1\}$ then an ordered sample of size r from a population of size n may be thought of as a point in the product space $\mathfrak{B}_1 \times \mathfrak{B}_2 \times \dots \times \mathfrak{B}_r$. In the specification of a random sample, the choice of an element does not influence the choice of the next element which is selected uniformly from the remaining subpopulation. Consequently, we may view a random sample of size r from a population of size n as obtained via sequential independent sampling from the experiments corresponding to the sample spaces $\mathfrak{B}_1, \mathfrak{B}_2, \dots, \mathfrak{B}_r$ in turn. The probability that a given ordered sample is obtained is hence given, in the “falling factorial”

notation introduced in Section II.5, by $\frac{1}{n} \cdot \frac{1}{n-1} \cdots \frac{1}{n-r+1} = \frac{1}{n^r}$, as it should.

As an ordered sample of size n from a population of size n is just a permutation of the elements, the procedure that we have outlined shows that a random permutation of n elements may be considered to be generated by independent sampling from the sequence of experiments $\mathfrak{B}_1, \mathfrak{B}_2, \dots, \mathfrak{B}_n$. The reader should contrast this product space with that of the previous example.

This example shows that it may be possible to represent the result of a probability experiment as an element of two different product spaces. Sometimes one representation or the other will turn out to be mathematically more convenient or amenable to analysis in applications.

5) *Random graphs.* A graph is a collection of vertices $\{1, \dots, n\}$, together with a collection of unordered pairs of vertices $\{i, j\}$ called edges. It can be represented schematically by representing vertices as points in a figure and connecting those pairs of vertices that are the edges of the graph with lines. Thus, there are eight possible graphs on three vertices, one graph with no edges, three with one edge, three with two edges, and one with three. See Figure 4. In general, there are

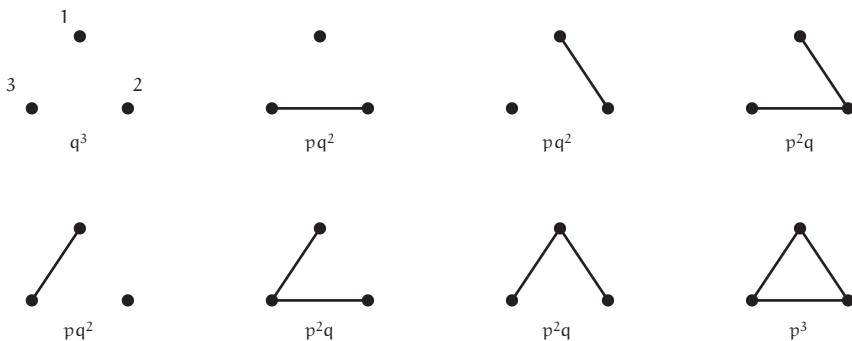


Figure 4: Random graphs on three vertices and their probabilities.

$2^{\binom{n}{2}}$ possible graphs on n vertices [as there are exactly $\binom{n}{2}$ unordered pairs of vertices $\{i, j\}$ and each of these is either included or not included in the graph].

Graphs are ubiquitous in many branches of mathematics, science, and engineering. To give a (very) small sample of applications, a wireless communication network may be represented for many purposes as a graph with users as vertices and edges between users who are within communication distance of each other; in chemical process control, a graph may indicate various stages of a process as vertices with edges connecting stages that are causally related; graphs play a major rôle in the design and analysis of algorithms in computer science with vertices representing decision points and edges leading to other decision points; the genetic composition of the nematode *C. elegans* is

best represented as a graph; neural connectivity in brain tissue has a natural graph structure; and, of course, everyone is familiar with family trees.

In graph applications where the underlying topology (that is to say, the collection of edges) of the graph is not strongly regulated, useful information can be garnered by considering the properties of random instances of the graph.

The *random graph* $G_{n,p}$ is a graph on n vertices selected as follows: for each pair of vertices i and j , the edge $\{i, j\}$ is included in the graph with probability p (and excluded with probability $q = 1 - p$) independently of all other pairs of vertices. Thus, the graph $G_{n,p}$ may be viewed as the outcome of a probability experiment: the sample space is the space of $2^{\binom{n}{2}}$ graphs on n vertices and the probability measure assigns to any graph with k edges the probability $p^k q^{\binom{n}{2}-k}$. The probabilities associated with the various graphs on three vertices, for example, are indicated in Figure 4. These graphs were introduced in a monumental paper of Paul Erdős and Alfred Rényi in 1960.²

There is a natural product space associated with the graph $G_{n,p}$. Order the $\binom{n}{2}$ pairs of vertices $\{i, j\}$ in some order, say, lexicographic. Now toss a coin with success probability p and failure probability $q = 1 - p$ repeatedly a total of $\binom{n}{2}$ times. For $1 \leq k \leq \binom{n}{2}$, include the k th edge in the graph if the k th toss was successful and exclude it from the graph if the toss was a failure. If we represent success by 1 and failure by 0, then the k th toss takes a value ω_k in the discrete space $\mathfrak{A} = \{1, 0\}$ and the conjoined experiment engenders the product space $\Omega = \mathfrak{A}^n$ as the $\binom{n}{2}$ -fold product of \mathfrak{A} with itself. The graph $G_{n,p}$ may be identified with a sample point $\omega = (\omega_1, \omega_2, \dots, \omega_{\binom{n}{2}})$ which tabulates the exact sequence of successes and failures obtained in the combined experiment engendered by repeated independent trials.

An alternative product space is occasionally preferable. Instead of lexicographically dealing with the edges, the idea is to group edges by vertex taking care not to double count. Associate with vertex i all undirected pairs $\{i, j\}$ with $j > i$. Thus, with vertex 1 is associated the $n - 1$ pairs $\{1, 2\}, \{1, 3\}, \dots, \{1, n\}$, with vertex 2 the $n - 2$ pairs $\{2, 3\}, \{2, 4\}, \dots, \{2, n\}$, and so on, with vertex $n - 1$ the single pair $\{n - 1, n\}$. With these associations, each potential edge is counted exactly once. We now construct a sequence of discrete sample spaces, $\mathfrak{B}_1, \mathfrak{B}_2, \dots, \mathfrak{B}_{n-1}$ as follows. The elements of \mathfrak{B}_i are $(n - i)$ -tuples $\omega_i = (\omega_{i,i+1}, \omega_{i,i+2}, \dots, \omega_{i,n})$ where $\omega_{ij} = 1$ if $\{i, j\}$ is an edge of $G_{n,p}$ and $\omega_{ij} = 0$ otherwise. As each $\{i, j\}$ is associated with one and only one vertex, the experiments corresponding to the \mathfrak{B}_i are independent and $G_{n,p}$ may be identified with an element $\omega = (\omega_1, \omega_2, \dots, \omega_{n-1})$ of the product space $\mathfrak{B}_1 \times \mathfrak{B}_2 \times \dots \times \mathfrak{B}_{n-1}$.

²P. Erdős and A. Rényi, "On the evolution of random graphs", *Magyar Tud. Akad. Mat. Kut. Int. Közl.*, vol. 5, pp. 17–61, 1960.

Thus far we have dealt with finite product spaces. The next step is to consider infinite product spaces arising from an unending sequence of independent trials. Such spaces may be thought of as idealisations of actual (perhaps only approximately) observable product spaces. If intuition is to serve at all in such a setting, we would want the following assertion to hold: *if A_1, A_2, \dots are events with A_i determined exclusively by the i th trial then the events are independent.* Or, more specifically, for every finite subcollection of these events $A_{i_1}, A_{i_2}, \dots, A_{i_n}$, we have a product law $P(A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_n}) = P(A_{i_1}) P(A_{i_2}) \dots P(A_{i_n})$.

The increasing complexity of a denumerably infinite number of trials unfortunately brings with it technical complications—the infinite product space is not discrete any longer, as we have seen in Example I.7.7. It is now not at all clear how one would proceed to specify a probability measure and whether, indeed, such a measure would be unique, or if it would satisfy the product law for events.

The existence and uniqueness of such a measure was shown by A. N. Kolmogorov in 1933 in a very general and abstract setting. As Mark Kac has pointed out, however, unrestrained abstraction demands a price in that it tends to conceal whole application areas whose key features are passed off as accidental in the abstract point of view. In Chapter V we will hence focus on constructing a *model* for coin tossing that contains the features of interest. It will transpire that the model will yield a unique probability measure containing the key product property. And along the way we will discover a deep and exciting connection with the theory of numbers. Armed with more intuition (and some more theory), we will return to Kolmogorov's abstract general structure in Chapter XII.

5 Independent families, Dynkin's π - λ theorem

If A and B are independent events then so are $\{\emptyset, A, A^c, \Omega\}$ and $\{\emptyset, B, B^c, \Omega\}$ in the sense that any pair of events, one from each set, is independent. A natural generalisation:

DEFINITION A finite collection of event families $\{A_1, \dots, A_n\}$ is independent if, for every choice of $A_1 \in \mathcal{A}_1, \dots, A_n \in \mathcal{A}_n$, the events A_1, \dots, A_n are independent. A sequence of event families $\{A_n, n \geq 1\}$ is independent if every finite subcollection of event families is independent.

Bernstein's construction of Example 2.7 shows that even if event families \mathcal{A} and \mathcal{B} are independent, derived events in these two families need not be. As independence deals with conjunctions we may close this loophole by considering classes of sets that are closed under intersections.

A π -class \mathcal{P} is a family of sets closed under intersection, that is, if $A \in \mathcal{P}$ and $B \in \mathcal{P}$ then $A \cap B \in \mathcal{P}$. Beginning with a sequence of independent π -classes it is now natural to ask whether independence carries over to the sequence of σ -algebras derived from these classes. Now a σ -algebra in an abstract space Ω is a complicated beast and

in analytical arguments it is preferable to reduce matters to considerations of simpler objects. E. B. Dynkin's π - λ theorem is one of the most useful results of this type and reduces considerations to classes of sets satisfying a monotone property.

A λ -class \mathcal{L} is a family of subsets of Ω satisfying: (1) $\Omega \in \mathcal{L}$; (2) \mathcal{L} is closed under monotone unions, that is, if $\{A_n \in \mathcal{L}, n \geq 1\}$ is an increasing sequence of subsets, $A_n \subseteq A_{n+1}$ for each n , then $\bigcup_n A_n \in \mathcal{L}$; and (3) \mathcal{L} is closed under monotone set differences, that is, if $A_1 \in \mathcal{L}$ and $A_2 \in \mathcal{L}$ with $A_1 \subseteq A_2$, then $A_2 \setminus A_1 \in \mathcal{L}$. By selecting A_2 to be Ω it is clear that a λ -class \mathcal{L} is closed under complementation. Furthermore, if $\{A_n, n \geq 1\}$ is an increasing family of sets then $\{A_n^c, n \geq 1\}$ is a decreasing family of sets and as, by de Morgan's laws, $\bigcap_n A_n^c = (\bigcup_n A_n)^c$, a λ -class \mathcal{L} is closed under monotone intersections as well. Thus, every λ -class is an example of a monotone class of sets. It is clear that a σ -algebra is itself a λ -class but that by only requiring closure under monotone operations, λ -classes have a larger writ.

A recurring motif in such settings is the exploitation of the *smallest* class of sets with a given property. This is the key to the proof of the theorem that follows.

DYNKIN'S π - λ THEOREM *Let \mathcal{P} be a π -class and \mathcal{L} a λ -class. If $\mathcal{P} \subseteq \mathcal{L}$ then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

PROOF: The intersection of all λ -classes containing all the elements of \mathcal{P} is by arguments similar to that for the minimal σ -algebra in Section I.8 itself a λ -class. We may as well suppose that \mathcal{L} is this *smallest λ -class over \mathcal{P}* and we do so without further ado.

It will suffice now to show that \mathcal{L} is itself a σ -algebra containing each of the elements of \mathcal{P} (whence, *a fortiori*, we would have $\sigma(\mathcal{P}) \subseteq \mathcal{L}$). We only need to show that \mathcal{L} is closed under countable unions. The crux of the argument is to show that \mathcal{L} preserves the closure under intersection property of the π -class \mathcal{P} it covers. We start with two, apparently increasingly refined, subclasses of \mathcal{L} :

$$\begin{aligned}\mathcal{L}' &= \{A \in \mathcal{L} : A \cap B \in \mathcal{L} \text{ for every } B \in \mathcal{P}\}, \\ \mathcal{L}'' &= \{A \in \mathcal{L} : A \cap B \in \mathcal{L} \text{ for every } B \in \mathcal{L}\}.\end{aligned}$$

Begin with the subclass \mathcal{L}' . Fix any $B \in \mathcal{P}$. As \mathcal{P} is a π -class, if A is in \mathcal{P} then $A \cap B$ is in \mathcal{P} , hence also in \mathcal{L} . It follows by definition of \mathcal{L}' that A is contained in \mathcal{L}' and so $\mathcal{P} \subseteq \mathcal{L}'$. We show now that \mathcal{L}' is in fact a λ -class over \mathcal{P} . Again fix any $B \in \mathcal{P}$.

(1) As $\Omega \cap B = B$ is in \mathcal{P} , hence also in \mathcal{L} , for each B in \mathcal{P} , it follows that $\Omega \in \mathcal{L}'$.

(2) Suppose $\{A_n, n \geq 1\}$ is an increasing family of sets in \mathcal{L}' . By the distributive property, $(\bigcup_n A_n) \cap B = \bigcup_n (A_n \cap B)$. By definition of \mathcal{L}' , $A_n \cap B \in \mathcal{L}$ for every n , and as these sets are increasing it follows that the limit set $\bigcup_n (A_n \cap B)$ is also in \mathcal{L} . Thus, $\bigcup_n A_n \in \mathcal{L}'$ and \mathcal{L}' is closed under monotone unions.

(3) Finally, suppose $A_1 \subseteq A_2$ are two elements of \mathcal{L}' . We may write $(A_2 \setminus A_1) \cap B = (A_2 \cap B) \setminus (A_1 \cap B)$ (if the reader draws a Venn diagram all will be clear). But, by definition of \mathcal{L}' , $A_1 \cap B$ and $A_2 \cap B$ are both in \mathcal{L} and it is clear furthermore that $(A_1 \cap B) \subseteq (A_2 \cap B)$. As \mathcal{L} is closed under monotone set differences it follows that $A_2 \setminus A_1$ is in \mathcal{L}' . Thus, \mathcal{L}' is also closed under monotone set differences.

So \mathcal{L}' is a λ -class containing the elements of \mathcal{P} . It is patent on the one hand that $\mathcal{L}' \subseteq \mathcal{L}$ but on the other hand \mathcal{L} is the smallest λ -class over \mathcal{P} . We are hence left with no alternative but to conclude that $\mathcal{L}' = \mathcal{L}$.

We are now primed to consider the refined subclass \mathcal{L}'' . Fix any B in $\mathcal{L}' = \mathcal{L}$. If A is in \mathcal{P} then $A \cap B$ is in \mathcal{L} by definition of \mathcal{L}' . But this means that A lies in \mathcal{L}'' . And so

$\mathcal{P} \subseteq \mathcal{L}''$. We now fix any $B \in \mathcal{L}$ and argue as in (1), (2), and (3) above (with \mathcal{P} replaced by \mathcal{L} and \mathcal{L}' by \mathcal{L}'') to deduce that \mathcal{L}'' is also a λ -class containing the elements of \mathcal{P} . We conclude anew that $\mathcal{L}'' = \mathcal{L}$.

Thus, the minimal λ -class over \mathcal{P} is closed under intersection (it is a π -class), complements, and (by de Morgan's laws) unions. But closure under unions as well as monotone unions implies closure under countable unions. Indeed, suppose A_1, A_2, \dots is any countable sequence of sets in \mathcal{L} . For each n , set $B_n = A_1 \cup \dots \cup A_n$. By induction, B_n is in \mathcal{L} for each n and as $\{B_n, n \geq 1\}$ is an increasing sequence of sets, it follows that \mathcal{L} also contains the limit set $\bigcup_n B_n = \bigcup_n A_n$. Thus the minimal λ -class over \mathcal{P} is indeed a σ -algebra as advertised, hence contains the minimal σ -algebra $\sigma(\mathcal{P})$. ▶

Another useful monotone class theorem of this stripe is due to P. Halmos. It replaces the "destination" λ -class \mathcal{L} by a monotone class with less structure at the expense of replacing the "originating" π -class \mathcal{P} by a class with more structure; see Problem 23.

The π - λ theorem is the piece of the puzzle needed to translate independence from π -classes to σ -algebras derived from them.

THEOREM 2 Suppose $\{\mathcal{P}_n, n \geq 1\}$ is a sequence of independent π -classes. Then the sequence $\{\sigma(\mathcal{P}_n), n \geq 1\}$ of generated σ -algebras is also independent.

PROOF: Suppose $\{\mathcal{P}_1, \dots, \mathcal{P}_n\}$ is an independent collection of π -classes. This is equivalent to saying that if, for each j , A_j is an event in the class \mathcal{P}_j augmented by the addition of Ω (if it is not already present in \mathcal{P}_j) then

$$\mathbf{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbf{P}(A_1) \times \mathbf{P}(A_2) \times \dots \times \mathbf{P}(A_n). \quad (5.1)$$

If we select one or more of the A_j to be Ω then we see that the product relation holds for all subsets of a collection of n events as per Definition 1.3. As a π -class augmented by the addition of Ω remains closed under intersection we may as well suppose that each \mathcal{P}_j contains Ω .

Let \mathcal{L} be the subclass of those events A_1 in $\sigma(\mathcal{P}_1)$ for which (5.1) holds for every selection of $A_2 \in \mathcal{P}_2, \dots, A_n \in \mathcal{P}_n$. We argue that \mathcal{L} is indeed a λ -class of sets. It is clear that \mathcal{L} contains the sets of \mathcal{P}_1 , hence has Ω as an element. It will suffice hence to show that \mathcal{L} is closed under monotone unions and monotone set differences.

Fix any $A_2 \in \mathcal{P}_2, \dots, A_n \in \mathcal{P}_n$. Suppose $\{B_k, k \geq 1\}$ is an increasing sequence of sets with $B_k \in \mathcal{L}$ for each k . Write $B = \bigcup_k B_k$ for the limit set. By definition of \mathcal{L} , we have $\mathbf{P}(B_k \cap A_2 \cap \dots \cap A_n) = \mathbf{P}(B_k) \mathbf{P}(A_2) \dots \mathbf{P}(A_n)$. But $B_k \uparrow B$ and $B_k \cap A_2 \cap \dots \cap A_n \uparrow B \cap A_2 \cap \dots \cap A_n$ as $k \rightarrow \infty$. By continuity of probability measure it follows that $\mathbf{P}(B \cap A_2 \cap \dots \cap A_n) = \mathbf{P}(B) \mathbf{P}(A_2) \dots \mathbf{P}(A_n)$ so that the limit set B is also in \mathcal{L} . Thus, \mathcal{L} is closed under monotone unions.

Suppose now that $A \subseteq B$ are two sets in \mathcal{L} . As we may decompose $B = A \cup (B \setminus A)$ as a union of disjoint sets, by additivity we have

$$\begin{aligned} \mathbf{P}((B \setminus A) \cap A_2 \cap \dots \cap A_n) &= \mathbf{P}(B \cap A_2 \cap \dots \cap A_n) - \mathbf{P}(A \cap A_2 \cap \dots \cap A_n) \\ &= (\mathbf{P}(B) - \mathbf{P}(A)) \mathbf{P}(A_2) \dots \mathbf{P}(A_n) = \mathbf{P}(B \setminus A) \mathbf{P}(A_2) \dots \mathbf{P}(A_n), \end{aligned}$$

and it follows that $B \setminus A \in \mathcal{L}$. Thus, \mathcal{L} is indeed a λ -class over \mathcal{P}_1 . By the π - λ theorem, $\sigma(\mathcal{P}_1) \subseteq \mathcal{L}$, and, as all sets in \mathcal{L} have the product property (5.1), it follows that

$\{\sigma(\mathcal{P}_1), \mathcal{P}_2, \dots, \mathcal{P}_n\}$ is an independent collection of π -systems. We now repeat the argument with \mathcal{L} generated from $\sigma(\mathcal{P}_2)$ to conclude that $\{\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \mathcal{P}_3, \dots, \mathcal{P}_n\}$ is an independent collection of π -systems. Arguing inductively we are left with the conclusion that $\{\sigma(\mathcal{P}_1), \sigma(\mathcal{P}_2), \dots, \sigma(\mathcal{P}_n)\}$ is an independent collection of derived σ -algebras.

As independence for a countable collection is defined in terms of independence of finite collections, the general result follows. ▶

Thus, independence across π -systems is inherited by the derived sets in the σ -algebras generated by these systems as well. In rough terms, very much in accordance with intuition, a family of events is independent if each of these events is derived from non-overlapping collections of independent events. Very satisfactory.

6 Problems

1. *Mutually exclusive events.* Suppose A and B are events of positive probability. Show that if A and B are mutually exclusive then they are not independent.
2. *Conditional independence.* Show that the independence of A and B given C neither implies, nor is implied by, the independence of A and B .
3. If A , B , and C are independent events, show that $A \cup B$ and C are independent.
4. *Patterns.* A coin turns up heads with probability p and tails with probability $q = 1 - p$. If the coin is tossed repeatedly, what is the probability that the pattern $\mathfrak{T}, \mathfrak{H}, \mathfrak{H}, \mathfrak{H}$ occurs before the pattern $\mathfrak{H}, \mathfrak{H}, \mathfrak{H}, \mathfrak{H}$?
5. *Runs.* Let Q_n denote the probability that in n tosses of a fair coin no run of three consecutive heads appears. Show that

$$Q_n = \frac{1}{2} Q_{n-1} + \frac{1}{4} Q_{n-2} + \frac{1}{8} Q_{n-3}$$

with the recurrence valid for all $n \geq 3$. The boundary conditions are $Q_0 = Q_1 = Q_2 = 1$. Determine Q_8 via the recurrence. [Hint: Consider the first tail.]

6. *Equivalent conditions.* In some abstract probability space, consider three events A_1 , B_1 , and C_1 , with complements A_2 , B_2 , and C_2 , respectively. Prove that A_1 , B_1 , and C_1 are independent if, and only if, the eight equations

$$\mathbf{P}\{A_i \cap B_j \cap C_k\} = \mathbf{P}(A_i) \mathbf{P}(B_j) \mathbf{P}(C_k), \quad i, j, k \in \{1, 2\},$$

all hold. Does any subset of these equations imply the others? If so, determine a minimum subset with this property.

7. *Continuation.* Let A_1, \dots, A_n be events in some probability space. For each i , define the events $A_{i0} = A_i$ and $A_{i1} = A_i^c$. Show that the events A_1, \dots, A_n are independent if, and only if, for every sequence of values j_1, \dots, j_n in $\{0, 1\}$, we have $\mathbf{P}(\bigcap_{i=1}^n A_{ij_i}) = \prod_{i=1}^n \mathbf{P}(A_{ij_i})$.

8. *Random sets.* Suppose A and B are independently selected random subsets of $\Omega = \{1, \dots, n\}$ (not excluding the empty set \emptyset or Ω itself). Show that $\mathbf{P}(A \subseteq B) = \left(\frac{3}{4}\right)^n$.

9. *Continuation.* With the random sets A and B as in the previous problem, determine the probability that A and B are disjoint.

10. Non-identical trials. In a repeated sequence of independent trials, the k th trial produces a success with probability $1/(2k + 1)$. Determine the probability P_n that after n trials there have been an odd number of successes.

11. Permutations. Consider the sample space comprised of all $k!$ permutations of the symbols a_1, \dots, a_k , together with the sequences of k symbols obtained by repeating a given symbol k times. Assign to each permutation the probability $1/(k^2(k - 2)!)$ and to each repetition probability $1/k^2$. Let A_i be the event that the symbol a_1 occurs in the i th location. Determine if the events A_1, \dots, A_n are pairwise independent and whether they are (jointly) independent.

12. Independence over sets of prime cardinality. Suppose p is a prime number. The sample space is $\Omega = \{1, \dots, p\}$ equipped with probability measure $P(A) = \text{card}(A)/p$ for every subset A of Ω . (As usual, $\text{card}(A)$ stands for the cardinality or size of A .) Determine a necessary and sufficient condition for two sets A and B to be independent.

13. Tournaments. Players of equal skill are pitted against each other in a succession of knock-out rounds in a tournament. Starting from a pool of 2^n individuals, players are paired randomly in the first round of the tourney with losers being eliminated. In successive rounds, the victors from the previous rounds are randomly paired and the losers are again eliminated. The n th round constitutes the final with the victor being crowned champion. Suppose A is a given player in the pool. Determine the probability of the event A_k that A plays exactly k games in the tournament. Let B be another player. What is the probability of the event E that A plays B in the tournament?

14. Guessing games. A fair coin is tossed. If the outcome is heads then a bent coin whose probability of heads is $5/6$ (and probability of tails is $1/6$) is tossed. Call this a type I bent coin. If the outcome of the fair coin toss is tails then another bent coin whose probability of tails is $5/6$ (and probability of heads is $1/6$) is tossed. Call this a type II bent coin. A gambler is shown the result of the toss of the bent coin and guesses that the original toss of the fair coin had the same outcome as that of the bent coin. What is the probability the guess is wrong?

15. Continuation, majority rule. In the setting of the previous problem, suppose that, following the toss of the fair coin, n type I coins are tossed or n type II coins are tossed depending on whether the outcome of the fair coin toss was heads or tails, respectively. Here n is an odd positive integer. The gambler is shown the results of the n bent coin tosses and uses the majority of the n tosses to guess the outcome of the original fair coin toss. What is the probability that the guess is wrong?

16. Continuation, majority rule with tie breaks. In the setting of Problem 14, suppose that four type I or type II coins are tossed depending on the outcome of the fair coin toss. The gambler again uses a majority rule to guess the outcome of the fair coin toss but asks for four more tosses of the bent coin in the event of a tie. This continues until she can make a clear majority decision. What is the probability of error?

17. Noisy binary relays. A binary symmetric channel is a conduit for information which, given a bit $x \in \{0, 1\}$ for transmission, produces a copy $y \in \{0, 1\}$ which is equal to x with probability $1 - p$ and differs from x with probability p . A starting bit $x = x_0$ is relayed through a succession of independent binary symmetric channels to create a sequence of bits, $x_0 \mapsto x_1 \mapsto \dots \mapsto x_n \mapsto \dots$, with x_n the bit produced by the n th channel in response to x_{n-1} . Let P_n denote the probability that the bit x_n produced

after n stages of transmission is equal to the initial bit x . By conditioning on the n th stage of transmission set up a recurrence for P_n in terms of P_{n-1} and solve it. Verify your answer by a direct combinatorial argument.

18. Gilbert channels.³ A wireless communication link may be crudely modelled at a given point in time by saying that it is either in a good state 0 or in a bad state 1. A chance-driven process causes the channel to fluctuate between good and bad states with possible transitions occurring at epochs 1, 2, The channel has no memory and, at any transition epoch, its next state depends only on its current state: given that it is currently in state 0 it remains in state 0 with probability p and transits to state 1 with probability $q = 1 - p$; given that it is currently in state 1 it remains in state 1 with probability p' and transits to state 0 with probability $q' = 1 - p'$. Initially, the channel is in state 0 with probability α and in state 1 with probability $\beta = 1 - \alpha$. Let α_n be the probability that the channel is in state 0 at epoch n . Set up a recurrence for α_n and thence determine an explicit solution for it. Evaluate the long-time limit $\alpha := \lim_{n \rightarrow \infty} \alpha_n$ and interpret your result. Propose an urn model that captures the probabilistic features of this problem.

19. Recessive and lethal genes. In the language of Section 3, suppose that individuals of type aa cannot breed. This is a simplification of the setting where the gene a is recessive and lethal; for instance, if aa -individuals are born but cannot survive; or, perhaps, via the enforcement of sterilisation laws in a draconian, eugenically obsessed society borrowing from Spartan principles. Suppose that in generation n the genotypes AA , Aa , and aa occur in the frequencies $u_n : 2v_n : w_n$ in both male and female populations. Conditioned upon only genotypes AA and Aa being permitted to mate, the frequencies of the genes A and a in the mating population are given by $p_n = (u_n + v_n)/(1 - w_n)$ and $q_n = v_n/(1 - w_n)$, respectively. Assuming random mating determine the probabilities of the genotypes AA , Aa , and aa in generation $n + 1$. Assume that the actual frequencies $u_{n+1} : 2v_{n+1} : w_{n+1}$ of the three genotypes in generation $n + 1$ follow these calculated probabilities and thence show the validity of the recurrence $q_{n+1}^{-1} = 1 + q_n^{-1}$ for the evolution of the lethal gene a in the mating population of each generation. Show hence that the lethal gene dies out but very slowly.

20. Suppose that the genotypes AA , Aa , aa in a population have frequencies $u = p^2$, $2v = 2pq$, and $w = q^2$, respectively. Given that a man is of genotype Aa , what is the conditional probability that his brother is of the same genotype?

21. Consider the repeated throws of a die. Find a number n such that there is an even chance that a run of three consecutive sixes appears before a run of n consecutive non-six faces.

22. Consider repeated independent trials with three possible outcomes A, B, and C with probabilities p , q , and r , respectively. Determine the probability that a run of m consecutive A's will occur before a run of n consecutive B's.

23. Halmos's monotone class theorem. A monotone class \mathcal{M} is a family of sets which is closed under monotone unions and intersections: if $\{A_n, n \geq 1\}$ is an increasing sequence of sets in \mathcal{M} then $\bigcup_n A_n \in \mathcal{M}$; if $\{A_n, n \geq 1\}$ is a decreasing sequence of sets in \mathcal{M} then $\bigcap_n A_n \in \mathcal{M}$. Suppose \mathcal{A} is an algebra of events, \mathcal{M} a monotone class. If $\mathcal{A} \subseteq \mathcal{M}$ show that $\sigma(\mathcal{A}) \subseteq \mathcal{M}$. [Hint: Mimic the proof of Dynkin's $\pi-\lambda$ theorem.]

³Named after the communication theorist E. N. Gilbert who first examined these models for wireless communication in the presence of noise.

IV

Probability Sieves

Some of the earliest applications of probability dealt with interrelationships between events. Of particular interest were problems dealing with disjunctions (or unions) of events or, alternatively, conjunctions (or intersections) of events. Results of great power and subtlety can be obtained in these settings using elementary methods, the simple demonstrations sometimes lending the results an entirely spurious aura of triviality. While problems of this nature have a classical provenance they continue to be of great importance in the modern theory and new results keep reshaping the landscape, the combinatorial methods that we shall see in this chapter resurfacing with a vengeance.

C 1, 4, 6, 9
A 2, 3, 5, 7, 10
F 8, 11

1 Inclusion and exclusion

The probability of the conjunction of independent events is naturally expressed as a product of the constituent event probabilities. This leads to the hope that simple relations of this form may continue to hold even if there is a mild degree of dependence across events. In such situations it is useful to express the probabilities of unions of events in terms of their intersections.

Two events A, B: We begin with a consideration of the simplest case of the disjunction of two events. In simple cases such as this a Venn diagram like the cartoon shown in Figure 1 is useful for visualisation. One can now write down, almost by inspection, the relation

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

For a formal proof, we may write $A \cup B$ as the union of disjoint sets, for instance in the form $A \cup B = (A \setminus B) \cup B$. It follows by additivity of probability measure that $P(A \cup B) = P(A \setminus B) + P(B)$. But A is the union of the disjoint set $A \setminus B$ and $A \cap B$, whence $P(A \setminus B) = P(A) - P(A \cap B)$. The claimed result follows.

Three events A, B, C: The corresponding result for three events follows quickly

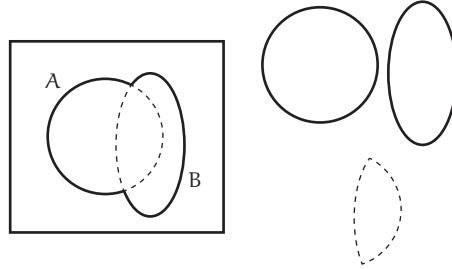


Figure 1: Event interrelations represented symbolically in a Venn diagram.

arguing along similar lines:

$$\begin{aligned} \mathbf{P}(A \cup B \cup C) &= \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) \\ &\quad - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C). \end{aligned}$$

A proof can be fashioned along the lines sketched above but we defer it as a more general result beckons. We can already see a pattern developing here: we include all singleton events, exclude all pairs, include all triplets, and so on. Hence the moniker, inclusion–exclusion.

n events A_1, \dots, A_n : The general result follows very much along the same pattern. It will be convenient first for $1 \leq k \leq n$ to define

$$S_k := \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k})$$

and to complete the specification for $k = 0$ by setting $S_0 = 1$. (The choice is dictated both by convenience and the natural convention that the intersection over an empty set is the entire sample space; of course, the union over an empty set is empty.)

THE THEOREM OF INCLUSION–EXCLUSION *The probability that one or more of the events A_i occur is given by*

$$\mathbf{P}(A_1 \cup \dots \cup A_n) = \sum_{k=1}^n (-1)^{k-1} S_k. \quad (1.1)$$

More generally, let $P(m)$ denote the probability that exactly m of the events A_1, \dots, A_n occur. Then

$$\begin{aligned} P(m) &= \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} S_{m+k} \\ &= S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} - \dots + (-1)^{n-m} \binom{n}{m} S_n \end{aligned} \quad (1.2)$$

for each $0 \leq m \leq n$.

Before diving into the proof it would be well to consider an illustrative application in a simple setting.

EXAMPLE 1) *The independent case—a first look at the binomial distribution.* Suppose n balls are distributed at random in v urns. What is the probability that the first urn is non-empty? Let A_i denote the event that the i th ball is placed in the first urn. Then the event of interest is $\bigcup_{i=1}^n A_i$ and the inclusion–exclusion formula (1.1) may immediately be applied. Each A_i has probability $1/v$ and as the events A_i ($1 \leq i \leq n$) are independent it follows that, for each k ,

$$S_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \binom{n}{k} \left(\frac{1}{v}\right)^k$$

as there are precisely $\binom{n}{k}$ ways of selecting k distinct events out of n . In consequence, the probability that the first urn is not empty is given by

$$\sum_{k=1}^n (-1)^{k-1} \binom{n}{k} \left(\frac{1}{v}\right)^k = 1 - \sum_{k=0}^n (-1)^k \binom{n}{k} \left(\frac{1}{v}\right)^k = 1 - \left(1 - \frac{1}{v}\right)^n.$$

Of course, the reader will have observed that in this case a direct demonstration is simple, almost trite,

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - P\left(\bigcap_{i=1}^n A_i^c\right) = 1 - \left(1 - \frac{1}{v}\right)^n,$$

but it is always useful to have independent confirmation.

More generally, the probability that the first urn contains exactly m balls is given by

$$\begin{aligned} \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} S_{m+k} &= \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} \binom{n}{m+k} \left(\frac{1}{v}\right)^{m+k} \\ &= \binom{n}{m} \left(\frac{1}{v}\right)^m \sum_{k=0}^{n-m} (-1)^k \binom{n-m}{k} \left(\frac{1}{v}\right)^k = \binom{n}{m} \left(\frac{1}{v}\right)^m \left(1 - \frac{1}{v}\right)^{n-m}. \end{aligned}$$

The reader may wish to amuse herself by providing a direct combinatorial argument.

The argument remains irreproachable if each A_i has probability p not necessarily rational. With a view to later expedience, we introduce the notation

$$b_n(m; p) := \binom{n}{m} p^m (1-p)^{n-m} \tag{1.3}$$

where m and n are non-negative integers, $m \leq n$, and $0 \leq p \leq 1$. The probability that at least one A_i occurs is then given by $1 - P(0) = 1 - b_n(0; p) = 1 - (1 - p)^n$ while the probability that exactly m of the A_i occur is given by $P(m) = b_n(m; p)$. The reader should verify that for each choice of n and p the numbers $b_n(m; p)$ ($0 \leq m \leq n$) form a discrete probability distribution, that is to say, $b_n(m; p) \geq 0$ for each m and $\sum_m b_n(m; p) = 1$: *this is the binomial distribution with parameters n and p .*

►

Before embarking on a proof of the inclusion-exclusion theorem we verify that $P(A_1 \cup \dots \cup A_n) = 1 - P(A_1^c \cap \dots \cap A_n^c) = 1 - P(0)$ so that (1.1) does indeed follow from (1.2). The form of (1.1) is suggestive, however, and it is worthwhile starting from the particular result first to spy out the lie of the land. At least two approaches are indicated.

A FIRST PROOF: A SAMPLE POINT ARGUMENT. A sample point-by-sample point approach to proof is both direct and general. Let E be any sample point. If E is not in $\bigcup_k A_k$ then there will be no contribution due to E to either the left or the right-hand side of (1.1). Now suppose E is an element of $\bigcup_k A_k$. Then E occurs in at least one of the A_k ; suppose that it occurs in *exactly* L of the events A_k . Then E contributes once to the probability on the left-hand side, while on the right, there can be a contribution due to E only for $k \leq L$ whence the total contribution due to E on the right is exactly

$$\sum_{k=1}^L (-1)^{k-1} \binom{L}{k} = 1 - \sum_{k=0}^L \binom{L}{k} (-1)^k = 1 - (1-1)^L = 1.$$

Thus, each sample point contributes exactly its own fair share to both left and right-hand sides.

We may argue for the general formulation (1.2) of the theorem of inclusion and exclusion along similar lines. Suppose E lies in exactly L of the A_i . If $L < m$ then clearly E contributes nothing to either side of (1.2). If $L = m$ then E contributes exactly once to the left-hand side and also haply once to the right-hand side (in the term S_m). If, finally $L > m$, then E again contributes nothing to $P(m)$; the number of times it contributes to the right-hand side of (1.2) is readily computed as

$$\sum_{k=0}^{L-m} (-1)^k \binom{m+k}{m} \binom{L}{m+k} = \binom{L}{m} \sum_{k=0}^{L-m} (-1)^k \binom{L-m}{k}.$$

To finish up, the binomial theorem shows that the right-hand side is given identically by $\binom{L}{m} (1-1)^{L-m} = 0$ so that when $L > m$ the point E contributes nothing to the right-hand side as well.

How does the proof generalise to continuous settings, and in general, to non-discrete spaces? For each i , let B_i be either the event A_i or its complement

A_i^c and consider the family of events E of the form $\bigcap_{i=1}^n B_i$. Observe that the sets in this family are disjoint: if $E^{(1)} = \bigcap_i B_i^{(1)}$ and $E^{(2)} = \bigcap_i B_i^{(2)}$ are any two distinct sets in this family then $E^{(1)} \cap E^{(2)} = \emptyset$. (Why? There is at least one index i for which $B_i^{(1)}$ and $B_i^{(2)}$ are complements.) Furthermore, this family of sets partitions each of the A_i , A_i^c , and all of Ω . (Why? Consider, for instance, the union of all the sets in this family for which $B_i = A_i$.) If one replaces the sample points E in the above argument by the events E in this family the argument carries through impeccably by additivity of probability measure. ►

Sample point arguments such as the one above are ubiquitous and powerful. The reader who is not convinced may find the alternative, equally useful, approach based on induction more appealing.

A SECOND PROOF: INDUCTION. Consider (1.1) anew. The base case has already been established for the case of two events. Now suppose (1.1) holds for all choices of n events. Write $B = A_1 \cup \dots \cup A_n$. The induction base applied to the disjunction $B \cup A_{n+1}$ quickly yields

$$\mathbf{P}(A_1 \cup \dots \cup A_n \cup A_{n+1}) = \mathbf{P}(B) + \mathbf{P}(A_{n+1}) - \mathbf{P}(B \cap A_{n+1}).$$

Now, by induction hypothesis, the probability of B is given by

$$\begin{aligned} \mathbf{P}(B) &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k}) \\ &= \mathbf{P}(A_1) + \dots + \mathbf{P}(A_n) + \sum_{k=2}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k}). \end{aligned}$$

Another application of the induction hypothesis to the union of the n events $A_1 \cap A_{n+1}, A_2 \cap A_{n+1}, \dots, A_n \cap A_{n+1}$ yields

$$\begin{aligned} \mathbf{P}(B \cap A_{n+1}) &= \sum_{k=1}^n (-1)^{k-1} \sum_{1 \leq i_1 < \dots < i_k \leq n} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k} \cap A_{n+1}) \\ &= - \sum_{k=2}^n (-1)^{k-1} \sum_{\substack{1 \leq i_1 < \dots < i_{k-1} \leq n \\ i_k = n+1}} \mathbf{P}(A_{i_1} \cap \dots \cap A_{i_k}) - (-1)^n \mathbf{P}(A_1 \cap \dots \cap A_{n+1}). \end{aligned}$$

Collecting terms shows that $\mathbf{P}(A_1 \cup \dots \cup A_n \cup A_{n+1}) = \sum_{k=1}^{n+1} (-1)^{k-1} S_k$ and this completes the induction. The reader should think on how to adapt this argument to demonstrate (1.2). ►

Inclusion-exclusion is most efficacious when the probabilities of conjunctions of events are easy to estimate. The independent case is the simplest and leads to the binomial distribution as we have seen. But the approach can pay dividends even when there are dependencies among the events and it is not clear how a direct approach may be fashioned.

EXAMPLE 2) Matchings, revisited. The method of inclusion and exclusion provides a new perspective on de Montmort's *le problème des rencontres* that we encountered in Section II.4. Stripped of local colour, the problem takes the form of matchings: let π_1, \dots, π_n be a random permutation of the numbers from 1 through n . We say that there is a *coincidence* at location k if $\pi_k = k$. What is the probability that there is at least one coincidence?

Let A_k connote the event that there is a coincidence at location k . Then $Q_n := P(A_1 \cup \dots \cup A_n)$ is the probability of at least one coincidence and the problem is tailor made for an application of inclusion–exclusion.

Of all permutations of the digits from 1 through n , the number of permutations in which any *given* set of k locations exhibit coincidences is $(n-k)!$ as the remaining $n-k$ numbers may be arbitrarily distributed. As all permutations of the digits from 1 through n are equally likely, it follows that the probability that any given set of locations i_1, \dots, i_k exhibit coincidences is given by

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = \frac{(n-k)!}{n!} \quad (1 \leq i_1 < \dots < i_k \leq n).$$

As there are exactly $\binom{n}{k}$ different ways of selecting k locations, that is to say, distinct selections for the k -set $\{i_1, \dots, i_k\}$, it follows that the quantity S_k in (1.1) is given by

$$S_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \frac{n!}{k!(n-k)!} \frac{(n-k)!}{n!} = \frac{1}{k!}.$$

Let $P_n(m)$ be the probability that there are exactly m coincidences. Inclusion–exclusion then immediately yields

$$P_n(m) = \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} S_{m+k} = \frac{1}{m!} \sum_{k=0}^{n-m} \frac{(-1)^k}{k!}. \quad (1.4)$$

It follows that, for every m , $P_n(m) \rightarrow e^{-1}/m!$ as $n \rightarrow \infty$ and, indeed, $P_n(m)$ is approximately given by its limiting value almost independently of the choice of n by virtue of the very rapid convergence of the exponential series $e^{-1} = \sum_{k=0}^{\infty} (-1)^k/k!$. In particular, we recover de Montmort's result

$$Q_n = 1 - P_n(0) = 1 - \sum_{k=0}^n \frac{(-1)^k}{k!} \approx 1 - e^{-1} = 0.63212\dots.$$

As we saw in Section II.4, the probability of achieving at least one matching is approximately two-thirds and is almost independent of n . The reader may find some profit in comparing approaches. ►

The inclusion–exclusion sums lead effortlessly to a variety of summation identities. Here is one.

EXAMPLE 3) *Sampling with replacement and a combinatorial identity.* Suppose that an ordered sample of size $m < n$ is selected with replacement from a population of size n . Let A_i be the event that the i th element of the population is not included in the sample. As there are $(n - 1)^m$ ordered samples which do not include i out of the total of n^m possible ordered samples, it follows that $P(A_i) = (n - 1)^m/n^m$. Likewise, for any choice of k distinct population elements i_1, \dots, i_k , we have $P(A_{i_1}, \dots, A_{i_k}) = (n - k)^m/n^m$ and

$$S_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \binom{n}{k} \frac{(n - k)^m}{n^m}.$$

By inclusion–exclusion, the probability that one or more elements of the population are excluded from the sample is hence given by

$$\sum_{k=1}^n (-1)^{k-1} \binom{n}{k} \frac{(n - k)^m}{n^m}.$$

If $m < n$ it is certain that at least one element of the population is excluded from the sample and the sum above equals one identically. It is a little more aesthetically pleasing to begin the summation at zero and by subtracting one and multiplying throughout by n^m we obtain the pleasant combinatorial identity

$$\sum_{k=0}^n (-1)^k \binom{n}{k} (n - k)^m = 0 \quad (1.5)$$

valid for all choices of positive integers m and n with $m < n$. ▶

The inclusion–exclusion identity has the metaphorical effect of sifting through the events until it finds one that occurs; interpreted thus, it can be viewed allegorically as a probability sieve, the oldest of a variety of sieve methods in probability. In spite of its elementary nature the basic inclusion–exclusion method has proved to be remarkably ubiquitous as the reader will discover in the following sections.

2 The sieve of Eratosthenes

The classical applications of inclusion–exclusion were in the theory of numbers. Write $\gcd(a, b)$ for the greatest common divisor of two natural numbers a and b , and for every real number x let $\lfloor x \rfloor$ denote the greatest integer $\leq x$.

THEOREM 1 *Let N be a natural number and a_1, \dots, a_n natural numbers that are relatively prime, that is to say, $\gcd(a_i, a_j) = 1$ if $i \neq j$. Let R be a random number*

selected from $\{1, \dots, N\}$. Then the probability that R is divisible by none of the a_i is given by

$$1 - \sum_{1 \leq i \leq n} \frac{1}{N} \left\lfloor \frac{N}{a_i} \right\rfloor + \sum_{1 \leq i < j \leq n} \frac{1}{N} \left\lfloor \frac{N}{a_i a_j} \right\rfloor - \dots + (-1)^n \frac{1}{N} \left\lfloor \frac{N}{a_1 a_2 \cdots a_n} \right\rfloor. \quad (2.1)$$

PROOF: Identify A_i as the event that a_i divides R . As the number of strictly positive integers $\leq N$ that are divisible by a_i is $\lfloor N/a_i \rfloor$, the term S_1 in the inclusion-exclusion formula is given by $\sum_{1 \leq i \leq n} \frac{1}{N} \lfloor \frac{N}{a_i} \rfloor$. As a_i and a_j are relatively prime if $i \neq j$, the number of strictly positive integers $\leq N$ that are divisible both by a_i and a_j is $\lfloor N/(a_i a_j) \rfloor$ and the term S_2 in the inclusion-exclusion formula is given by $\sum_{1 \leq i < j \leq n} \frac{1}{N} \lfloor \frac{N}{a_i a_j} \rfloor$. Proceeding in this fashion, the k th term is given by $S_k = \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} \frac{1}{N} \lfloor \frac{N}{a_{i_1} a_{i_2} \cdots a_{i_k}} \rfloor$. The probability that none of the a_i divide R is then given by $1 - S_1 + S_2 - \dots + (-1)^n S_n$. ▶

Multiplying the expression in (2.1) throughout by N yields the number of positive integers $\leq N$ that are relatively prime to each of a_1, \dots, a_n and the intrinsically combinatorial nature of the result is again manifest.

For each natural number N , Euler's totient function $\varphi(N)$ is defined to be the number of positive integers k , no larger than N , which are relatively prime to N . This function is of great importance in number theory.

THEOREM 2 For each positive integer $N \geq 1$, we have $\varphi(N) = N \prod_p (1 - \frac{1}{p})$ where the product extends over all prime divisors p of N .¹

PROOF: Suppose there are n prime divisors, p_1, \dots, p_n of N and in Theorem 1 identify the a_i with the p_i . As a positive integer k is relatively prime to N if, and only if, k is divisible by none of the prime divisors p_i of N , it follows that

$$\varphi(N) = N - \sum_{1 \leq i \leq n} \frac{N}{p_i} + \sum_{1 \leq i < j \leq n} \frac{N}{p_i p_j} - \dots + (-1)^n \frac{N}{p_1 p_2 \cdots p_n} \quad (2.2)$$

and the right-hand side is seen to be just the expansion of $N \prod_{i=1}^n (1 - \frac{1}{p_i})$. ▶

A compact expression for the totient function can be obtained in terms of another fundamental number-theoretic function, the Möbius function $\mu(N)$, defined on the natural numbers. The prime factorisation theorem asserts that every integer $N \geq 2$ may be written as a product of primes, $N = p_1 p_2 \cdots p_\nu$, where $\nu = \nu(N)$ is the number of prime factors in the product counting repetitions. The factorisation is unique up to rearrangements of the terms of the product. If a prime factor p_k is repeated in the factorisation then N is divisible by p_k^2 ; if, on the other hand, N has no repeated prime factors (so that each

¹The reader should bear in mind that the number 1 is not considered to be a prime.

p_k is distinct) then it is not divisible by the square of any prime. The Möbius function is defined on the natural numbers based on this characterisation. We begin by setting $\mu(1) = 1$. For $N \geq 2$, we set $\mu(N) = 0$ if N is divisible by a square. Finally, if N has no repeated prime factors, we set $\mu(N) = 1$ if it has an even number of distinct prime factors, and $\mu(N) = -1$ if it has an odd number of distinct prime factors. A quick look at the first few values of the Möbius function shows its highly irregular character as it skitters around the values $\{-1, 0, 1\}$. But it is just the ticket to capture the right-hand side of (2.2) and we may write the totient function in the form $\varphi(N) = N \sum_d \mu(d)/d$ where the sum is allowed to range over all positive divisors d of N . The Möbius function in the summand will summarily drop those terms corresponding to integers divisible by the square of any of the primes p_i , leaving only the terms with alternating sign in the inclusion–exclusion formula.

Expressions such as (2.2) require explicit knowledge of the primes $\leq N$ and luckily a procedure from antiquity known as the *sieve of Eratosthenes* yields an effective procedure for their systematic tabulation. Suppose one knows in advance all the primes $\leq \sqrt{N}$, say, q_1, \dots, q_m . Begin with the sequence of integers $2, 3, \dots, N$ and strike out all those integers that are divisible by $q_1 = 2$; of the remaining, eliminate those that are divisible by $q_2 = 3$; then eliminate those divisible by $q_3 = 5$, and proceed in this fashion until all the residual numbers divisible by q_m are eliminated. The numbers that remain are all strictly larger than \sqrt{N} but no larger than N . Moreover, none of the numbers arising out of the sieve of Eratosthenes is divisible by any of the primes q_i that are less than or equal to \sqrt{N} , nor expressible as a product of two numbers larger than \sqrt{N} . In other words, these numbers constitute all the primes that are $> \sqrt{N}$ and $\leq N$. Proceeding systematically in this fashion, one can determine all the primes $\leq N$ for any given value of N . The reader may prefer the ditty given alongside, while it is admittedly imprecise as a description of the sieve, to the long-winded description of the procedure.

How many primes are churned out by the sieve of Eratosthenes? In standard notation, write $\pi(x)$ for the number of primes less than or equal to any positive real number x . Then the number of primes exceeding \sqrt{N} but not larger than N is $\pi(N) - \pi(\sqrt{N})$, but this is given by Theorem 1 to be equal to

$$-1 + N - \sum_{1 \leq i \leq m} \left\lfloor \frac{N}{q_i} \right\rfloor + \sum_{1 \leq i < j \leq m} \left\lfloor \frac{N}{q_i q_j} \right\rfloor - \dots + (-1)^m \left\lfloor \frac{N}{q_1 q_2 \cdots q_m} \right\rfloor.$$

The sign of the terms in the sum alternates based upon whether the number of prime factors in the denominator is even or odd and so the expression may be written compactly in terms of the Möbius function: *the number of primes exceeding \sqrt{N} but not larger than N is given by $\pi(N) - \pi(\sqrt{N}) = -1 + \sum_d \mu(d)[N/d]$*

where the sum ranges over all positive divisors d of the product $q_1 q_2 \cdots q_m$. The factor -1 arises because the positive integer 1 is not divisible by any of q_1, \dots, q_m and is hence also accounted for in the count given by (2.1) with the a_i identified with the primes q_i that are $\leq \sqrt{N}$.

As an aside, the *prime number theorem* asserts that $\pi(x)/\frac{x}{\log x} \rightarrow 1$ as $x \rightarrow \infty$. This fundamental result in the theory of numbers was proved independently by Hadamard and de la Vallée Poussin in 1896.

3 On trees and a formula of Cayley

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a mathematical object consisting of an abstract set \mathcal{V} of objects called *vertices* together with a collection \mathcal{E} of unordered pairs of vertices called *edges*. We may view a graph pictorially by representing vertices by points on a sheet of paper and drawing lines between points to represent edges. In a vivid language we may then say that an edge $\{i, j\}$ is *incident on* the vertices i and j and say colloquially that i and j are *neighbours* or are *adjacent*. The *degree* of a vertex i , denoted $\deg(i)$, is the number of edges incident upon it, that is to say, the number of its neighbours.

A graph is *connected* if, starting from any vertex, we can move to any other vertex by traversing only along the edges of the graph. Formally, \mathcal{G} is connected if, for every pair of distinct vertices u and v , there exists $k \geq 0$ and a sequence of distinct vertices $u = i_0, i_1, \dots, i_k, i_{k+1} = v$ so that $\{i_j, i_{j+1}\}$ is an edge of the graph for $0 \leq j \leq k$. Thus, we may traverse from u to v through the sequence of vertices $u = i_0 \rightarrow i_1 \rightarrow \cdots \rightarrow i_k \rightarrow i_{k+1} = v$ by progressing along the edges between succeeding vertices; we call such a progression a *path*. Thus, a graph is connected if, and only if, there is a path between any two vertices. An example of a (large) connected graph is the internet with users, computers, and routers representing vertices, and edges between entities that are directly linked.

Connected graphs can have fantastically complicated structures and a principle of parsimony may suggest that we begin with a consideration of the simplest kind of graph structure that is still connected. Suppose the graph contains a sequence of edges $\{i_0, i_1\}, \{i_1, i_2\}, \dots, \{i_{k-1}, i_k\}, \{i_k, i_0\}$ where none of the vertices i_1, \dots, i_k repeat. Starting at vertex i_0 and traversing these edges of the graph in sequence returns us to i_0 . Naturally enough, we will call such a sequence of edges a *cycle*. Now it is clear that if \mathcal{G} is not already cycle-free, then we may remove any edge (but not the associated vertices) from any cycle of \mathcal{G} without affecting connectivity (though the paths which originally involved the eliminated edge will now become longer). This pruning process has created a new connected graph with one fewer edge. Repeated iterations of the pruning procedure will terminate in a finite number of steps (as there are only a finite number of possible edges) in a cycle-free connected subgraph of \mathcal{G} on the orig-

inal set of vertices (that is to say, a connected graph on the n vertices which contains no cycles and whose edges are all also edges of the parent graph \mathcal{G}).

This leads to a consideration of cycle-free connected graphs, or *trees*. A pictorial view of such graphs (see Figure 2) suggests an arborisation of the edges of these objects, hence the name. As trees are connected graphs, every

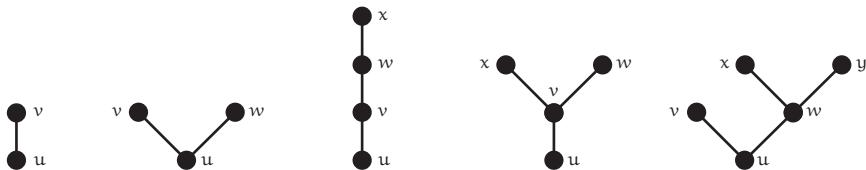


Figure 2: Trees on two, three, four, and five vertices. New trees may be formed by permutations of the vertices.

vertex has degree at least one. Vertices of degree exactly one in the tree have a special character as they can only appear as terminating vertices in paths; no vertex of degree one can be an intermediate vertex of a path. The arboreal image is irresistible and we call vertices of degree one *leaves* of the tree.

It is clear from our construction that every connected graph on n vertices contains an imbedded tree (possibly many). The importance of trees in applications is that they represent a minimal subgraph at which the body politic remains connected. In a communications network, for instance, information flows from vertex to vertex along the edges of the graph. In such settings it occasionally becomes important to propagate some (systems-level) information to all vertices in the network. A naïve approach to broadcasting the information through the network begins with the vertex which is the source of the information transmitting it to each of its neighbours along its edges. Each neighbouring vertex in turn broadcasts the information to all its neighbours. The procedure repeats and ultimately, as the graph is connected, all vertices are privy to the information. The procedure as outlined is wasteful as any vertex on one or more cycles will receive multiple copies of the same piece of information. If, alternatively, we first identify an imbedded tree in the graph and only transmit along the edges of the tree then much wasteful communication is saved. How much? We begin with a preliminary characterisation.

THEOREM 1 *Every finite tree contains at least two leaves.*

PROOF: We proceed by contradiction. Suppose a tree on n vertices contains no leaves. Then every vertex has degree ≥ 2 . Beginning with any vertex i_0 we proceed along an edge to a neighbouring vertex i_1 . As i_1 has degree at least two we may then proceed along a hitherto unused edge to a new vertex i_2 . Again, i_2 has degree at least two so that an unused edge incident on i_2 is available leading to a vertex i_3 . Now i_3 cannot be coincident with i_0 as otherwise a cycle would

be formed. Using an unused edge incident on i_3 we now proceed to a new vertex i_4 not previously encountered and, proceeding in this fashion, construct an infinite sequence of neighbouring vertices $i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_r \rightarrow \dots$ where each vertex is distinct. But this is impossible as we only have a finite number of vertices in the graph. Indeed, a vertex must repeat after no more than $n - 1$ steps leading to a cycle. Thus, a tree must have at least one leaf.

Suppose now that a tree on n vertices has exactly one leaf, say, i_0 . Starting with i_0 we move along its solitary edge to its neighbour i_1 . As i_1 has degree ≥ 2 we may now move along an unused edge to a new vertex i_2 . We can now repeat the previous argument and, as every new vertex we encounter has degree ≥ 2 , construct an infinite sequence of neighbouring vertices $i_0 \rightarrow i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_r \rightarrow \dots$ where each vertex is distinct. But this is again impossible. It follows that every finite tree must have at least two vertices of degree one. ▶

If $n > 2$ it is clear that no two leaves in a tree can be neighbours as they would otherwise be isolated. It follows that, if $n > 2$, there must exist at least one vertex with degree ≥ 2 . Thus, a tree on $n > 2$ vertices must have at least $n/2$ edges. The absence of cycles, on the other hand, does not permit the number of edges of the tree to get too large.

THEOREM 2 *Every tree on n vertices contains $n - 1$ edges.*

PROOF: The case $n = 2$ follows by inspection and we suppose accordingly that $n > 2$. To set up a recursive argument, write $n_1 = n$ and suppose the tree has k_1 leaves. (We now know that $n_1 > k_1 \geq 2$.) As these vertices can only terminate paths, by excising them and their associated edges from the tree we must be left with a tree on the remaining $n_2 = n_1 - k_1$ vertices. If $n_2 > 2$ we may repeat the procedure. Suppose this pruned tree has k_2 leaves. Removing these and the associated edges in turn results in a tree on $n_3 = n_2 - k_2$ vertices. Pruning this tree in the suggested manner leads to a tree on n_4 vertices, and so on. As n_1, n_2, n_3, \dots is a strictly decreasing sequence, proceeding in this fashion, after a finite number of stages, say, $r - 1$, we will be left with a tree on n_{r-1} vertices where either $n_{r-1} > 2$ and it has $k_r = n_{r-1} - 1$ leaves, or $n_{r-1} = 2$, in which case we set $k_r = 1$ and designate one of the two vertices as a leaf. In either case we prune the $k_r = n_{r-1} - 1$ designated vertices and the associated edges leaving a trivial tree on a single vertex, $n_r = 1$. The number of edges that have been pruned is $k_1 + \dots + k_r$ and as no edges are left this must be the number of edges in the tree. But as each edge is removed with an attendant vertex, this sum must also coincide with the number of vertices that have been removed. As only one vertex remains standing, it follows that $k_1 + \dots + k_r = n - 1$. ▶

As the number of edges in a connected graph on n vertices is potentially as large as $\binom{n}{2}$, globally disseminating information from a source node only

along the edges of an imbedded tree instead of broadcasting over all edges can potentially save a substantial amount of communication.

Let \mathcal{T}_n be the family of trees on n vertices. What can be said about the number of trees, $T_n = \text{card } \mathcal{T}_n$? A randomisation argument provides a beautiful answer.

For $n = 1$ we have only the trivial tree consisting of one vertex, whence $T_1 = 1$; for $n = 2$ we have a solitary tree consisting of two vertices connected by an edge, whence $T_2 = 1$ also. We now consider the cases $n > 2$. Suppose that a tree is chosen randomly from \mathcal{T}_n , each tree having equal probability $1/T_n$ of selection. Let A_i denote the event that a given vertex i is a leaf of the random tree. Now i will be a leaf of the tree if, and only if, there is precisely one edge connecting it to one of the remaining $n - 1$ vertices and, as this edge cannot be on a path connecting two other vertices, the pruned graph with vertex i and its associated edge removed forms a tree on the remaining $n - 1$ vertices. It follows that $P(A_i) = (n - 1)T_{n-1}/T_n$.

Building on this argument, suppose i_1, \dots, i_k are distinct vertices. Each of these vertices will be a leaf of a tree in \mathcal{T}_n if, and only if, each of the vertices i_j has as neighbour a vertex in $\{1, \dots, n\} \setminus \{i_1, \dots, i_k\}$ and the graph obtained by excising the vertices i_1, \dots, i_k and the associated edges forms a tree on the remaining $n - k$ vertices. Accordingly, $P(A_{i_1}, \dots, A_{i_k}) = (n - k)^k T_{n-k}/T_n$, and we begin to discern the beginnings of a sieve argument. The inclusion-exclusion summands are given by

$$S_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \binom{n}{k} (n - k)^k \frac{T_{n-k}}{T_n} \quad (1 \leq k \leq n),$$

so that the probability that a random tree in \mathcal{T}_n has at least one leaf is given by

$$\sum_{k=1}^n (-1)^{k-1} \binom{n}{k} (n - k)^k \frac{T_{n-k}}{T_n}.$$

But every tree must have at least one leaf and so this probability must be identically equal to one. We hence obtain the elegant recurrence

$$T_n = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} (n - k)^k T_{n-k} \quad (n \geq 3).$$

Running the recurrence through a few steps shows that

$$T_3 = 6T_2 - 3T_1 = 3,$$

$$T_4 = 12T_3 - 24T_2 + 4T_1 = 16,$$

$$T_5 = 20T_4 - 90T_3 + 80T_2 - 5T_1 = 125,$$

$$T_6 = 30T_5 - 240T_4 + 540T_3 - 240T_2 + 6T_1 = 1296,$$

and it is not too hard to see the emerging pattern.

CAYLEY'S FORMULA $T_n = n^{n-2}$.

PROOF: As induction hypothesis, suppose that $T_j = j^{j-2}$ for $j \leq n - 1$. (The base cases $j = 1$ and $j = 2$ trivially satisfy the induction hypothesis as $T_1 = 1^{1-2} = 1$ and $T_2 = 2^{2-2} = 1$.) Then

$$\begin{aligned} \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} (n-k)^k T_{n-k} &= \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} (n-k)^{n-2} \\ &= n^{n-2} - \sum_{k=0}^n (-1)^k \binom{n}{k} (n-k)^{n-2}. \end{aligned}$$

Setting $m = n - 2$ in the combinatorial identity (1.5) shows that the sum on the right is identically zero and this completes the induction. ▶

The formula was discovered by C. W. Borchardt in 1860 and extended by A. Cayley (to whom it is now generally attributed) in 1889. While a large number of proofs are now known for Cayley's formula, few can match the beauty and elegance of the sieve argument due to J. W. Moon that I have given here.²

4 Boole's inequality, the Borel–Cantelli lemmas

The monotone property of probability measure yields the natural inequality $\mathbf{P}(A) \geq \mathbf{P}(B)$ whenever $A \supseteq B$. The method of inclusion and exclusion allows us to determine a whole slew of results along these lines.

The first and simplest inequality is that of Boole. It crops up in an incredibly diverse range of applications, far out of proportion to its apparently trivial nature. Begin with the observation that we can write $A \cup B$ as a disjoint union via $A \cup B = (A \setminus B) \cup B$. And as $A \setminus B$ is palpably contained in A , it follows by monotonicity of probability measure that

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

A repeated inductive application yields the basic result known as *Boole's inequality*: for every n , and every choice of events A_1, \dots, A_n ,

$$\mathbf{P}(A_1 \cup A_2 \cup \dots \cup A_n) \leq \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots + \mathbf{P}(A_n).$$

For obvious reasons, Boole's inequality is sometimes also referred to as the *union bound* in the literature.

We can extend Boole's inequality to a countable union of sets by exploiting the continuity property of probability measure.

²J. W. Moon, "Another proof of Cayley's formula for counting trees", *The American Mathematical Monthly*, vol. 70, no. 8, pp. 846–847, 1963.

THEOREM 1 Suppose $\{A_k, k \geq 1\}$ is any sequence of events in some probability space. Then $P(\bigcup_k A_k) \leq \sum_k P(A_k)$.

PROOF: For finite sequences the result is just a statement of Boole's inequality. Consider now the case of a denumerably infinite sequence of events $\{A_k, k \geq 1\}$. For each n , write $C_n = \bigcup_{k=1}^n A_k$. It is clear that $\{C_n\}$ is an increasing sequence of events, $C_n \subseteq C_{n+1}$ for each n , whence $C_n = \bigcup_{m=1}^n C_m = \bigcup_{k=1}^n A_k$, and $\bigcup_{n=1}^{\infty} C_n = \bigcup_{k=1}^{\infty} A_k$. The continuity axiom of probability measure (see also Problem I.28) now asserts that

$$\lim_{n \rightarrow \infty} P(C_n) = P\left(\bigcup_{n=1}^{\infty} C_n\right) = P\left(\bigcup_{k=1}^{\infty} A_k\right).$$

On the other hand, Boole's inequality directly yields $P(C_n) \leq \sum_{k=1}^n P(A_k)$ for each n . Taking the limit as $n \rightarrow \infty$ of both sides of the inequality, we obtain $\lim_{n \rightarrow \infty} P(C_n) \leq \sum_{k=1}^{\infty} P(A_k)$ as asserted. ▶

A very simple application of subadditivity can frequently be deployed to great effect. First some terminology.

Let $\{A_n, n \geq 1\}$ be any sequence of events in a probability space. For each n , let $B_m = \bigcup_{n \geq m} A_n$ be the union of the sets A_m, A_{m+1}, \dots . It is clear that $\{B_m, m \geq 1\}$ is a decreasing sequence of events and we write $\limsup_n A_n = \bigcap_m B_m$ for the limiting event. The terminology is intended to mirror the idea of the limit superior of a real sequence (see Section XXI.1 in the Appendix).

An analogous procedure leads to the limit inferior of a sequence of sets. Form the increasing sequence of sets $C_m = \bigcap_{n \geq m} A_n$ and denote $\liminf_n A_n = \bigcup_m C_m$. As $\liminf_n A_n = (\limsup_n A_n^c)^c$ the notation is redundant but it is convenient to have both forms at hand just as it is useful to have separate symbols for addition and subtraction.

Now it is clear that a sample point ω lies in $\limsup_n A_n$ if, and only if, it lies in each of the sets B_m . If this is the case then ω must lie in an infinity of the sets A_n . Indeed, suppose to the contrary that ω lies only in a finite number of the A_n . Then there exists a number M such that $\omega \notin A_n$ for $n \geq M$. But then the sets B_M, B_{M+1}, \dots cannot contain ω . It follows that *the event $\limsup_n A_n$ occurs if, and only if, A_n occurs infinitely often, in short, A_n i.o.* This evocative terminology was introduced by Kai Lai Chung. Subadditivity provides a very simple test of when $\limsup_n A_n$ occurs.

THEOREM 2 Suppose the series $\sum_n P(A_n)$ converges. Then $P(\limsup_n A_n) = P\{A_n \text{ i.o.}\} = 0$ or, in words, with probability one only finitely many of the events A_n occur.

PROOF: Writing $B_m = \bigcup_{n \geq m} A_n$, we have $P(B_m) \leq \sum_{n=m}^{\infty} P(A_n)$ by subadditivity of probability measure. But the upper bound must tend to zero

as $m \rightarrow \infty$ because the assumed convergence of the series $\sum_{n=1}^{\infty} P(A_n)$ implies that the tails must vanish. The claimed result follows by continuity as $\{B_m, m \geq 1\}$ decreases to $\limsup_n A_n$. ▶

One of the great virtues of a theorem of this stripe is the absence of complex preconditions on the constituent events. The reader will be struck by the absence of any requirement on independence or, indeed, any structure at all among the A_n : all that is needed is that the event probabilities decrease rapidly enough that their sum is convergent.

A satisfactory converse to the theorem is known only in the case when there is an independent skeleton. The classical, though not the strongest, form of the converse is given below. We begin with an elementary inequality that the reader may well be familiar with.

LEMMA *The inequality $1 + x \leq e^x$ holds for all x .*

PROOF: If the reader sketches the graphs of the functions on either side of the inequality she will see that the claimed result is evident, the graphs also suggesting the simple proof. As $e^0 = 1$ and $\frac{d}{dx}e^x = e^x$, the mean value theorem tells us that $e^x = 1 + xe^{\xi}$ for some $\xi = \xi(x)$ between 0 and x . As $e^{\xi} \geq 1$ for $x \geq 0$ and $0 < e^{\xi} \leq 1$ for $x < 0$, we have $x \leq xe^{\xi}$ for all x and we conclude that $1 + x \leq e^x$. ▶

THEOREM 3 *Suppose the series $\sum_n P(A_n)$ diverges. If the events $\{A_n, n \geq 1\}$ are independent then $P(\limsup_n A_n) = P\{A_n \text{ i.o.}\} = 1$ and the events A_n occur infinitely often with probability one.*

PROOF: Write $p_n = P(A_n)$. If $\{A_n, n \geq 1\}$ is independent then so is $\{A_n^c, n \geq 1\}$ and so, by applying the elementary exponential inequality to $x = -p$, we have

$$\begin{aligned} P(A_m^c \cap A_{m+1}^c \cap \cdots \cap A_{m+N}^c) &= (1 - p_m)(1 - p_{m+1}) \cdots (1 - p_{m+N}) \\ &\leq \exp(-(p_m + p_{m+1} + \cdots + p_{m+N})) \end{aligned}$$

for every $m \geq 1$ and $N \geq 1$. As the series $\sum_n p_n$ is divergent we have $\sum_{n=m}^{m+N} p_n \rightarrow \infty$ as $N \rightarrow \infty$ for every choice of m . Consequently, $P(\bigcap_{n=m}^{m+N} A_n^c) \rightarrow 0$ as $N \rightarrow \infty$. By continuity of probability measure (see Problem I.29) it follows that $P(\bigcap_{n=m}^{\infty} A_n^c) = 0$ or, equivalently, $P(\bigcup_{n=m}^{\infty} A_n) = 1$, for each m , and hence $P\{A_n \text{ i.o.}\} = 1$. ▶

The strong independence requirement makes this theorem much less useful than its preceding counterpart.

Theorems 2 and 3 are called the *Borel–Cantelli lemmas*. As we shall see, the simple subadditive condition espoused in Theorem 2 provides the final piece of the puzzle in the strong law of large numbers.

The inequality of Boole and, more generally, countable subadditivity can already be deployed in elementary settings, frequently with devastating effect. In the next section we turn immediately to a classical historical example in the theory of numbers; additional examples of the use of Boole's inequality in divers contexts will be found scattered through subsequent chapters.

5 Applications in Ramsey theory

Will an arbitrary group of 7 members of a diffuse social network necessarily contain a subgroup of 3 friends or a subgroup of 3 strangers? The answer, perhaps surprisingly, is “Yes”. Indeed, any group of 6 people will contain a subgroup of 3 mutual acquaintances or a subgroup of 3 mutual strangers. A group of 5 individuals, however, does not necessarily have this property as the reader may verify from Figure 3: solid lines indicate mutual acquaintance, dashed lines indicate the two parties are strangers. The reader may wish to try her hand at showing that such an example cannot be contrived with 6 individuals no matter how acquaintanceship is assigned between pairs.

This sounds like a party game, amusing but of no great moment, but turns out to be the starting point of a very rich and profitable combinatorial investigation initiated by the English mathematician Frank Plumpton Ramsey before his untimely death in 1930 at the age of 26. Let us put the party game on a more formal footing.

The *complete graph* K_n is a graph on n vertices, which we may identify simply as $\mathcal{V} = \{1, 2, \dots, n\}$, for which every unordered pair of vertices $\{i, j\}$ is an edge, $\mathcal{E} = \{\{i, j\}, i \neq j\}$. A *two-colouring* of the edges of K_n is a map $\chi: \mathcal{E} \rightarrow \{\text{Red, Green}\}$ which maps to each edge $\{i, j\}$ one of the two colours Red or Green. (Of course, we could equally well have chosen 0 and 1 as the two “colours” instead of Red and Green; or solid and dashed lines, as in the adjoining figure. But the problem is more colourful this way.)

Our party problem can now be posed as follows: does K_n always contain a monochromatic K_k , i.e., a Red K_k or a Green K_k , for any two-colouring? (Our problem had $n = 7$ and $k = 2$.) Ramsey answered this, and more besides, in his celebrated theorem: *for any $k \geq 2$, there is a finite value of n for which any two-colouring of K_n contains a monochromatic K_k and so also there is a smallest n for which this is true.* The smallest such n is called the *Ramsey number* (associated with the pair (k, k)) and denoted $R(k, k)$.

The reader may perhaps worry that the definition seems to implicitly assume that the Ramsey numbers are finite. Is it possible that they are infinite? That is to say, is it possible that, for any n , there exists a two-colouring of the edges of K_n which does not contain an embedded monochromatic K_k ? Ramsey’s basic theorem showed that this is not the case. It is actually simpler to show this in a slightly more general context. For integers $j, k \geq 1$ we define the Ramsey number $R(j, k)$ to be the smallest value of n for which any two-colouring of the edges of K_n into the colours Red and Green contains a Red K_j or a Green K_k .

RAMSEY’S THEOREM *The numbers $R(j, k)$ are finite for each choice of integers $j, k \geq 1$.*

PROOF: A convenient base for an induction is established by adopting the convention

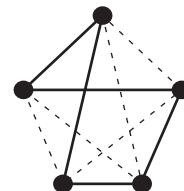


Figure 3: An acquaintance graph.

that a graph on a single vertex is monochromatic (the set of edges is vacuous and may be assigned any colour). Then $R(j, 1) = R(1, k) = 1$ for all j and k . We now establish by induction a finite upper bound for $R(j, k)$. As induction hypothesis suppose that $R(j - 1, k)$ and $R(j, k - 1)$ are finite. Consider now any two-colouring of the edges of a graph \mathcal{G} on $n = R(j - 1, k) + R(j, k - 1)$ vertices. Let v be any vertex of the graph. We partition the remaining $n - 1$ vertices into two groups: the collection \mathcal{U} of vertices u for which the edge $\{u, v\}$ is Red and the collection \mathcal{W} of vertices w for which the edge $\{u, w\}$ is Green. Then we must have $\text{card}(\mathcal{U}) + \text{card}(\mathcal{W}) = n - 1 = R(j - 1, k) + R(j, k - 1) - 1$. It follows that $\text{card}(\mathcal{U}) \geq R(j - 1, k)$ or $\text{card}(\mathcal{W}) \geq R(j, k - 1)$ as, if both inequalities are violated, we will be forced into the contradiction $\text{card}(\mathcal{U}) + \text{card}(\mathcal{W}) + 1 < n$. Suppose that $\text{card}(\mathcal{U}) \geq R(j - 1, k)$. If \mathcal{U} contains a Green K_k , so does the graph \mathcal{G} on n vertices and we are done. Else \mathcal{U} contains a Red K_{j-1} . But then these $j - 1$ vertices together with the vertex v form a Red K_j . An entirely similar argument shows that if, instead, $\text{card}(\mathcal{W}) \geq R(j, k - 1)$ then \mathcal{G} will contain a Red K_j or a Green K_k . It follows that any two-colouring of the edges of a graph on $n = R(j - 1, k) + R(j, k - 1)$ vertices will contain a Red K_j or a Green K_k and, hence, $R(j, k) \leq R(j - 1, k) + R(j, k - 1)$. ▶

The method of proof also suggests how to generate upper bounds on the rate of growth of these functions. The reader should be able to see how to extend the proof, with the natural extension of notation, to show the finiteness of the Ramsey numbers $R(k_1, \dots, k_c)$ for graph c -colourings. More general versions can also be proved.

Ramsey's theorem assures us that $R(k, k)$ is finite for each k . But what can we actually say about them? Table 1 details what is currently known about the first few Ramsey numbers. Remarkably, notwithstanding a great deal of scrutiny, the precise value of $R(5, 5)$ [and, *a fortiori*, also $R(6, 6)$] is still unknown (as of 2012). And the problem gets exponentially harder as k increases.

How hard is it to determine the Ramsey numbers? In this context, a well-known anecdote featuring Paul Erdős is worth repeating. Erdős asks us to imagine an alien race whose technology is vastly superior to us landing on earth and demanding that we produce the value of $R(5, 5)$; if we are in default of their demand they will destroy the planet. In this case Erdős recommends that we marshall the planet's scientific and computational resources together and in a desperate race against the clock attempt to determine $R(5, 5)$ to stave off extinction. If the aliens demand the value of $R(6, 6)$, however, Erdős believed that the better strategy would be to preemptively attempt to destroy the aliens, however futile the attempt may seem in the face of overwhelming force.

Can we say anything about how quickly these numbers grow with k ? In a landmark paper in 1947 that foreshadowed the explosion of activity in what is now called the *probabilistic method*, Erdős showed how a lower bound may be obtained almost effortlessly for these formidable numbers using a probabilistic

$R(2, 2) = 2,$
$R(3, 3) = 6,$
$R(4, 4) = 18,$
$43 \leq R(5, 5) \leq 49,$
$102 \leq R(6, 6) \leq 165.$

Table 1: Ramsey numbers.

argument.³ Here it is.

Let χ be a *random* two-colouring of K_n . The sample space here is the space of all two colourings. How many colourings are there? Well, each edge $\{i, j\}$ can be coloured in one of two ways and there are a total of $n(n-1)/2 = \binom{n}{2}$ distinct edges. It follows that the sample space contains $2^{\binom{n}{2}}$ points (colourings). By a random colouring we mean that each colouring χ has equal probability $2^{-\binom{n}{2}}$. Equivalently, we may consider that the random colouring is generated by flipping a fair coin repeatedly to determine the colour of each edge; thus, each edge has probability $1/2$ apiece of being Red or Green, edge colours determined independently of each other.

Let S be any fixed k -set of vertices in K_n and let A_S be the event that S forms a monochromatic K_k . The probability that the $\binom{k}{2}$ edges of S are all Red is, by independence of choice of edge colours, equal to $2^{-\binom{k}{2}}$, and likewise also for the probability that the edges of S are all Green. It follows that the probability that S is monochromatic is given by $P(A_S) = 2^{1-\binom{k}{2}}$. The event $\bigcup_{S: \text{card}(S)=k} A_S$ (where the union is over all $\binom{n}{k}$ choices of subsets S consisting of exactly k vertices) occurs if the subgraph induced on *any* k -set S is monochromatic. The probability of this disjunction of events is then exactly the probability that K_n with a random colouring χ contains a monochromatic K_k . The reader will quickly realise that an exact computation of the probability is difficult with dependencies rife across the events A_S as each edge of the graph is represented in $\binom{n-2}{k-2}$ of the k -sets A_S . Boole's inequality makes light of the dependencies, however, by quickly bounding the probability of interest from above via

$$P\left(\bigcup_{S: \text{card}(S)=k} A_S\right) \leq \sum_{S: \text{card}(S)=k} P(A_S) = \binom{n}{k} 2^{1-\binom{k}{2}}.$$

If $n \geq R(k, k)$ then K_n will contain a monochromatic K_k for every colouring of K_n , which is the same as saying $P(\bigcup A_S) = 1$. Conversely, if $n < R(k, k)$ then there exists at least one colouring of K_n for which one cannot exhibit even one monochromatic K_k , or, what is the same thing, $P(\bigcup A_S) < 1$. And thus we have retraced Erdős's path in 1947.

THEOREM 2 *For any positive integer k let $r = r(k)$ be the largest integer n satisfying the inequality $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$. Then $R(k, k) > r(k)$.*

The key step in the analysis was the deployment of the placid inequality of Boole to bound the probability of the disjunction of a large number of events with a complicated dependency structure.

³P. Erdős, "Some remarks on the theory of graphs", *Bulletin of the American Mathematics Society*, vol. 53, pp. 292–294, 1947.

How does this lower bound on $R(k, k)$ grow with k ? Observe that

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} < \frac{n^k}{k!}. \quad (5.1)$$

To obtain a lower bound on the factorial, pass to logarithms and, as the logarithm function increases monotonically, estimate the resulting sum by a corresponding integral which is readily evaluated by parts:

$$\log k! = \sum_{j=2}^k \log j > \int_1^k \log x \, dx = k \log k - (k-1) > k \log k - k. \quad (5.2)$$

It follows that $k! = e^{\log k!} > k^k e^{-k}$. [Stirling's formula gives better bounds: see (XIV.7.3) and (XVI.6.1).] Putting all the pieces together we obtain

$$P\left(\bigcup_{S:|S|=k} A_S\right) \leq \binom{n}{k} 2^{1-\binom{k}{2}} < n^k k^{-k} e^k 2^{1-\binom{k}{2}}.$$

For any $k \geq 2$, the upper bound on the right is less than one if $n \leq k 2^{k/2} / (2e\sqrt{2})$. This immediately provides a lower bound for the Ramsey numbers.

COROLLARY *For every positive integer k , the Ramsey number $R(k, k)$ may be bounded below by $R(k, k) > k 2^{k/2} / (2e\sqrt{2})$.*

We can improve matters slightly asymptotically. Stirling's formula tells us that $k! \sim \sqrt{2\pi k} k^k e^{-k}$ as $k \rightarrow \infty$. (The asymptotic "tilde" notation means that the ratio of the two sides tends to one.) As $k^{1/k} \rightarrow 1$ asymptotically, we obtain the improvement $R(k, k) > (1+\epsilon)k 2^{k/2} / (e\sqrt{2})$ for every choice of $\epsilon > 0$ provided k is sufficiently large. We can achieve a further marginal improvement in the lower bound by getting rid of the factor $\sqrt{2} e$ in the denominator as well via a slightly more sophisticated argument. Unfortunately this is all we can do and the result is still very far from known upper bounds for $R(k, k)$ of the order of $4^k/\sqrt{k}$. But, in the words of the probabilist J. Spencer, we do what we can.

Ramsey's investigations into the famous numbers bearing his name engendered a rich theory dealing with sequences of numbers. Here is another problem from the canon.

Suppose each of the numbers from 1 to 8 is coloured either black or white. Is it always the case that there is a monochromatic arithmetic sequence of length 3 embedded in any two-coloured sequence of the 8 numbers? [The reader will recall that an arithmetic sequence of length k has the form $x, x+a, x+2a, \dots, x+(k-1)a$.] The answer in this case is "No", the following sequence witness to the assertion.

● ○ ○ ● ● ○ ○ ●

The reader may wish to play with two-colourings of the numbers from 1 to 9 to examine whether the answer is now "Yes".

More generally, let k be a fixed positive integer. The *van der Waerden number* $W(k)$ is the smallest integer n such that *every* two-colouring of the numbers from 1 to n will have an embedded monochromatic arithmetic sequence of length k . Van der Waerden's celebrated theorem asserts that these numbers exist. The determination of these numbers is a long-standing open problem; Table 2 shows how shockingly little we know about them (*circa* 2012). In his original paper, van der Waerden actually provides an upper bound for these numbers that has the growth rate of the Ackermann functions—a ludicrously large rate of growth. Can anything be said in the reverse direction? The probabilistic method accords a quick and dirty bound.

We consider a random two-colouring of the numbers from 1 to n . Suppose S is any k -term arithmetic sequence contained within 1 to n and let A_S denote the event that S is monochromatic. Then $P(A_S) = 2^{1-k}$. We now observe that any length- k arithmetic sequence is completely specified by its first two terms. Consequently, the number of k -term arithmetic sequences contained in 1 to n is bounded above by $\binom{n}{2} < n^2/2$. Boole's inequality mops up:

$$P\left(\bigcup_S A_S\right) \leq \sum_S 2^{1-k} < \frac{1}{2} n^2 2^{1-k} = n^2 2^{-k}.$$

If the expression on the right is less than 1 then there is at least one two-colouring of 1 to n that does not have an imbedded monochromatic arithmetic subsequence of length k ; we have effortlessly obtained a lower bound for $W(k)$.

THEOREM 3 *For every $k \geq 1$, we have $W(k) > 2^{k/2}$.*

Sadly, this is very far from the upper bound; while we will improve the result shortly, the best known results are still far from satisfactory.

6 Bonferroni's inequalities, Poisson approximation

More sophisticated versions of Boole's inequality can be generated by systematically milking the alternating sign in inclusion–exclusion to generate an increasingly fine probabilistic sieve. These inequalities are a central part of what is referred to as the *Poisson paradigm*.

As before, let A_1, \dots, A_n be events of interest and let $P(m)$ denote the probability of the event that exactly m of the A_i occur. We recall that the inclusion–exclusion formula gives us the explicit result

$$P(m) = \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} S_{m+k}$$

k	$W(k)$
1	1
2	3
3	9
4	35
5	178
6	1,132
7	> 3,703
8	> 11,495

Table 2: Van der Waerden numbers.

where S_k is as before the sum of all the probabilities of conjunctions of the events A_i , k at a time. The alternating signs of the summands suggest that it may be profitable to examine the truncated sums

$$\sigma_K(m) = \sum_{k=0}^{K-1} (-1)^k \binom{m+k}{m} S_{m+k}. \quad (6.1)$$

By including all the terms in the sum it is clear that $\sigma_{n-m+1}(m) = P(m)$, but much more can be said.

THEOREM 1 *The partial sums $\sigma_K(m)$ bound $P(m)$ from below when K is even and from above when K is odd. More precisely, the approximation error $P(m) - \sigma_K(m)$ has the sign $(-1)^K$ of the first neglected term and is bounded in absolute value by that term.*

PROOF: We consider the family of events of the form $E = B_1 \cap \dots \cap B_n$ where each B_i is either A_i or A_i^c . This family of events partitions the sample space Ω and *a fortiori* partitions each A_i and A_i^c . Each event E in this family is hence either contained in any given A_i or is disjoint from it, i.e., is contained in A_i^c . In other words, $E \cap A_i$ is either E or \emptyset . We may suppose that any given event E of this form is contained in exactly $m + L$ of the sets A_i where $-m \leq L = L(E) \leq n - m$. As

$$P(m) - \sigma_K(m) = \sum_{k \geq K} (-1)^k \binom{m+k}{m} S_{m+k}, \quad (6.2)$$

the contribution of E to the sum on the right is identically 0 if $L < K$ (as $S_{m+k} = 0$ if $k > L$). If $L \geq K$ on the other hand, the contribution of E to the sum is given by

$$\begin{aligned} P(E) \sum_{k \geq K} (-1)^k \binom{m+k}{m} \binom{m+L}{m+k} &= P(E) \binom{m+L}{m} \sum_{k \geq K} (-1)^k \binom{L}{k} \\ &= P(E) \binom{m+L}{m} \sum_{k \geq K} (-1)^k \left[\binom{L-1}{k} + \binom{L-1}{k-1} \right], \end{aligned}$$

the last step following by an application of Pascal's triangle. The sum on the right telescopes with terms cancelling pairwise leaving only the term $(-1)^K \binom{L-1}{K-1}$. The contribution of E on the right of (6.2) is hence $(-1)^K \binom{m+L}{m} \binom{L-1}{K-1} P(E)$ which, it may be observed, is less in absolute value than $\binom{m+L}{m} \binom{L}{K} P(E)$, the absolute value of the contribution of E to the first term that has been left out of the truncated series. As each event E contributes either 0 or a term of sign $(-1)^K$ to the right-hand side of (6.2), it follows that $P(m) - \sigma_K(m) = (-1)^K C_K$ where C_K is a positive constant satisfying $C_K \leq \binom{m+K}{m} S_{m+k}$. ▶

By induction it follows quickly that we have established the hierarchy of inequalities known as *Bonferroni's inequalities*:

$$\sigma_2(m) \leq \sigma_4(m) \leq \cdots \leq \underbrace{\sigma_{n-m+1}(m)}_{=P(m)} \leq \cdots \leq \sigma_3(m) \leq \sigma_1(m),$$

The bookend inequalities $\sigma_2(m) \leq P(m) \leq \sigma_1(m)$ provide quick bounds for back-of-the-envelope calculations; the reader will recognise in the upper bound the inequality of Boole and, as we have seen, even such an apparently loose bound can be very useful. But Bonferroni's inequalities allow us to deduce much more: *for every even K*,

$$\sum_{k=0}^{K-1} (-1)^k \binom{m+k}{m} S_{m+k} \leq P(m) \leq \sum_{k=0}^K (-1)^k \binom{m+k}{m} S_{m+k}. \quad (6.3)$$

While the power of Bonferroni's inequalities is best appreciated after the development of a little more machinery, a classical example may serve to illustrate their utility.

EXAMPLE 1) The coupon collector's problem. Facing flagging sales a novelty supplier in London attempts to capitalise on a feeling of nostalgia current in the paying public by enclosing a copy of an original playbill of a randomly selected Gilbert and Sullivan opera produced in the years 1875–1896 with each piece of merchandise. The ploy is rewarded as a sentimental public flocks to the merchandise to try to obtain a complete set of copies of the playbills. Assuming that the playbill distributed with each novelty item is selected randomly and uniformly from the 13 operas produced in that period, how many items will an avid collector have to purchase before she can be assured of obtaining a complete set of playbills?

More generally, suppose a collector wishes to obtain a full set of n coupons (playbills in our example) and obtains a random coupon with each purchase. If she makes t purchases, what is the probability that she has not succeeded in obtaining a complete set of coupons? Write A_i for the event that the i th coupon has not been obtained after t purchases. These are the “bad” events. The “good” event of interest is $\bigcap_{i=1}^n A_i^c$, the event that all n coupons are obtained in t trials. As $P(\bigcap_{i=1}^n A_i^c) = 1 - P(\bigcup_{i=1}^n A_i)$ it suffices to estimate the probability that one or more of the coupons is not obtained in t trials. The stage is set for inclusion and exclusion.

It is clear that the probability of the event A_i is $(n-1)^t/n^t = (1-1/n)^t$. More generally, the probability that none of any given group of k events, say, A_{i_1}, \dots, A_{i_k} , is obtained in t trials is $(n-k)^t/n^t = (1-k/n)^t$. By inclusion and exclusion we now obtain

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k-1} \binom{n}{k} \left(1 - \frac{k}{n}\right)^t = 1 - \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t,$$

or, equivalently,

$$P\left(\bigcap_{i=1}^n A_i^c\right) = \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t.$$

Unfortunately, here the matter would appear to rest: the sum above cannot be simplified further in general and while it is nice to have an explicit solution, in itself the sum is not particularly informative.

Perhaps we can move forward if we are willing to settle for an asymptotic estimate for large values of n instead. Recalling that $(1 - c/x)^x$ is approximately e^{-c} for large values of x we may be tempted to carry out the following series of approximations:

$$\begin{aligned} \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t &\stackrel{?}{\approx} \sum_{k=0}^n \binom{n}{k} (-e^{-t/n})^k \\ &= (1 - e^{-t/n})^n \stackrel{?}{\approx} e^{-ne^{-t/n}}. \end{aligned}$$

If t is chosen such that $ne^{-t/n} = \lambda$, i.e., $t = n \log n - n \log \lambda$, then we have the speculative approximation $e^{-\lambda}$ for the probability that our dedicated collector manages to secure all n coupons. For small λ then, the probability of securing all the coupons is close to 1 while for large λ the probability is very small. And, if our approximations haven't broken down by this point, this would mean that the critical number of trials t is around $n \log n$. A critical look at our approximations suggests that we may be able to justify them if k does not become too large. An examination of the inclusion-exclusion formula truncated after a fixed number of terms is hence indicated. If we can establish that most of the contribution to the formula comes from the first few terms we will be home free.

Let K be any fixed positive integer and consider the truncated sum $\sum_{k=0}^K (-1)^k \binom{n}{k} (1 - k/n)^t$. As k varies in a fixed range the binomial coefficient $\binom{n}{k} = n^k/k!$ may be asymptotically estimated by $n^k/k!$ as $n \rightarrow \infty$. Indeed,

$$\lim_{n \rightarrow \infty} \binom{n}{k} \Big/ \frac{n^k}{k!} = \lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) = 1 \quad (6.4)$$

for each fixed k . We express this asymptotic equivalence succinctly by writing $\binom{n}{k} \sim n^k/k!$ as $n \rightarrow \infty$. To estimate the remaining summand, we recall the Taylor series for the logarithm,

$$\log(1 - x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \dots,$$

convergent whenever $|x| < 1$. It follows that

$$(1 - \frac{k}{n})^t = \exp\left\{t \log\left(1 - \frac{k}{n}\right)\right\} = \exp\left\{-\frac{tk}{n} - \frac{tk^2}{2n^2} - \dots\right\}$$

which may be estimated by $e^{-tk/n}$ if t is not too large. More precisely, suppose k takes values in a fixed range and $t = t_n$ varies with n such that $t/n^2 \rightarrow 0$ as $n \rightarrow \infty$. Then $(1 - k/n)^t \sim e^{-tk/n}$ asymptotically as $n \rightarrow \infty$ where, as before, we interpret the asymptotic relation \sim to mean that the ratio of the two sides tends to one as n tends to infinity. With these estimates in hand, we obtain

$$\sum_{k=0}^K (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t \sim \sum_{k=0}^K \frac{(-ne^{-t/n})^k}{k!} \quad (n \rightarrow \infty).$$

If $\lambda = ne^{-t/n}$ is bounded then the right-hand side is just a truncated Taylor series for the exponential $e^{-\lambda} = \sum_k (-\lambda)^k/k!$. As this series converges absolutely and uniformly over every bounded interval, it follows that, for every choice of $\epsilon > 0$, we can find a sufficiently large selection of $K = K(\epsilon)$ such that $|\sum_{k=0}^K (-\lambda)^k/k! - e^{-\lambda}| < \epsilon$. Bonferroni's inequalities now allow us to sandwich the desired probability above and below: for a sufficiently large choice of K the lower bound in the inclusion-exclusion formula exceeds $e^{-\lambda} - \epsilon$ while the upper bound is less than $e^{-\lambda} + \epsilon$. It follows that for every $\epsilon > 0$,

$$e^{-\lambda} - \epsilon < \sum_{k=0}^n (-1)^k \binom{n}{k} \left(1 - \frac{k}{n}\right)^t < e^{-\lambda} + \epsilon$$

for all sufficiently large n . As ϵ may be chosen arbitrarily small, it follows that the probability that all n coupons are collected tends to $e^{-\lambda}$ as $n \rightarrow \infty$ if the number of trials satisfies $t = \lceil n \log(n/\lambda) \rceil$.

Thus, in particular, if a collector of Gilbert and Sullivan playbills purchases 63 novelty items she will have collected all 13 playbills with a confidence level of 90%. If she desires a confidence level of 99% she will have to purchase 93 novelty items. ▶

An examination of the analysis of the coupon collector's problem shows that we have indeed proved a bit more. In the inclusion-exclusion formula (1.2), let us write $S_k = S_k^{(n)}$ and $P(m) = P_n(m)$ to explicitly acknowledge the dependence on the hidden parameter n . Bonferroni's inequalities simplify asymptotically into an elegant limit law, occasionally referred to as *Brun's sieve*.

THEOREM 2 *Let λ be any fixed positive quantity, m any fixed positive integer. If, for each fixed $k \geq 1$, $S_k^{(n)} \rightarrow \lambda^k/k!$ as $n \rightarrow \infty$, then $P_n(m) \rightarrow e^{-\lambda}\lambda^m/m!$ asymptotically with n .*

PROOF: Let K be any fixed but arbitrary positive integer. Rewriting (6.1) to explicitly acknowledge the dependence on n , we have

$$\sigma_K^{(n)}(m) = \sum_{k=0}^{K-1} (-1)^k \binom{m+k}{m} S_{m+k}^{(n)}.$$

As $n \rightarrow \infty$, the right-hand side tends to the limiting value

$$\sum_{k=0}^{K-1} (-1)^k \binom{m+k}{m} \frac{\lambda^{m+k}}{(m+k)!} = \frac{\lambda^m}{m!} \sum_{k=0}^{K-1} \frac{(-\lambda)^k}{k!}.$$

For any fixed $\epsilon > 0$, a sufficiently large selection of $K = K(\epsilon)$ ensures that the term on the right differs from $p(m; \lambda) := e^{-\lambda} \lambda^m / m!$ by less than ϵ in absolute value. Bonferroni's inequalities (6.3) now show that, for every $\epsilon > 0$ and all sufficiently large values of n ,

$$p(m; \lambda) - \epsilon < P_n(m) < p(m; \lambda) + \epsilon$$

or, what is the same thing, $P_n(m) \rightarrow p(m; \lambda)$ for every positive integer m . ▶

The reader should verify that the quantities $p(m; \lambda) = e^{-\lambda} \lambda^m / m!$ determine a discrete probability distribution on the positive integers $m \geq 0$: *this is the Poisson distribution with parameter λ .*

A special case of our theorem was discovered by S. D. Poisson in 1837. Suppose that, for each n , we have a family of independent events A_1, \dots, A_n each of common probability $P(A_i) = \pi_n$ where we allow these probabilities to depend on n . Then, following Example 1.1, the probability $P_n(m)$ that exactly m of the A_i occur is given by the binomial probability

$$b_n(m; \pi_n) = \binom{n}{m} \pi_n^m (1 - \pi_n)^{n-m}.$$

If the sequence $\{\pi_n, n \geq 1\}$ asymptotically satisfies $\lim_n n\pi_n = \lambda$ for some fixed $\lambda > 0$ then $S_1^{(n)} = n\pi_n \rightarrow \lambda$ and, in general, $S_k^{(n)} = \binom{n}{k} \pi_n^k \rightarrow \lambda^k / k!$ for each fixed k . We have rediscovered Poisson's famous approximation to the binomial which, in view of its historical significance as well as its practical importance, is worth enshrining.

THEOREM 3 *Suppose $\lambda > 0$ is a fixed constant and $\{\pi_n, n \geq 1\}$ is a sequence of positive numbers, $0 < \pi_n < 1$, satisfying $n\pi_n \rightarrow \lambda$. Then $b_n(m; \pi_n) \rightarrow p(m; \lambda)$ for each fixed positive integer m .*

A limit theorem such as this applies formally not to a fixed collection of events but, rather, to a whole sequence of event collections indexed by a parameter n . Its practical sway is seen in cases when, in a given application with a fixed event family, n is large and π_n is small. In such cases the theorem provides theoretical cover for the approximation of $b_n(m; \pi_n)$ by $p(m; \lambda)$ (though the error in approximation will, of course, depend on how well the asymptotic conditions are satisfied, to wit, the extent that $n\pi_n = \lambda$ is not too large). The reader will find the nearest applications in Section VIII.6.

Revisiting the earlier examples in view of the Poisson paradigm yields a more precise picture when event probabilities are small.

EXAMPLES: 2) *Balls, urns, and the binomial distribution.* We consider n balls randomly distributed in n urns. Let A_i be the event that the i th ball is placed in one of the first r urns. Arguing as in Example 1.1, the probability $P(m) = P_n(m; r)$ that the first r urns contain exactly m balls is given by

$$P_n(m; r) = \binom{n}{m} \left(\frac{r}{n}\right)^m \left(1 - \frac{r}{n}\right)^{n-m} = b_n(m; r/n).$$

It follows via the Poisson paradigm that $P_n(m; r) \rightarrow p(m; r)$ as $n \rightarrow \infty$ for every fixed r and m .

3) *The hat check problem, revisited.* In Example 1.2, $S_k^{(n)} = 1/k!$ for each k whence $P_n(m) \rightarrow e^{-1}/m!$ as $n \rightarrow \infty$. It follows that the number of sailors who retrieve their own hats is asymptotically governed by $p(m; 1)$, the Poisson distribution with parameter 1.

4) *The coupon collector's problem, revisited.* With the choice $t = n \log(n/\lambda)$ in the coupon collector's problem,

$$S_k^{(n)} = \binom{n}{k} \left(1 - \frac{k}{n}\right)^t \sim \frac{n^k}{k!} e^{-k \log(n/\lambda)} = \frac{\lambda^k}{k!}.$$

It follows that the number of coupons collected is governed asymptotically by $p(m; \lambda)$, the Poisson distribution with parameter λ . ▶

As we've seen, the Poisson distribution is intimately related to the binomial. Just as the binomial appears naturally when events forebode "typical" situations, the Poisson makes an appearance when we are interested in combinations of a large number of rare events; this is the province of extremal value theory.

7 Applications in random graphs, isolation

Construct a graph on n vertices as follows: for each distinct pair of vertices i and j , include the edge $\{i, j\}$ in the graph if a coin toss with success probability p is successful. There are $\binom{n}{2}$ coin tosses in all, one for each distinct (unordered) pair of vertices, and we suppose that the various coin tosses are all mutually independent. The result of the experiment is the *random graph* $G_{n,p}$ considered in Example III.4.5.

Random graphs were popularised by Paul Erdős and wide-ranging investigations have shown an amazing spectrum of properties. In a typical setting the number n of vertices is large and $p = p_n$ is a function (typically decreasing) of n . One is then interested in whether the graph exhibits such and such a property asymptotically.

To illustrate, say that a vertex i of the graph is *isolated* if there are no edges in $G_{n,p}$ that are incident on i , or more verbosely, $G_{n,p}$ contains none of the $n - 1$ edges $\{i, 1\}, \dots, \{i, i-1\}, \{i, i+1\}, \dots, \{i, n\}$. What is the probability that $G_{n,p}$ contains one or more isolated vertices? If p is large we anticipate that $G_{n,p}$ is unlikely to contain isolated vertices while if p is small then $G_{n,p}$ is likely to contain isolated vertices. What is the critical range $p = p_n$ at which isolated vertices first appear?

Let A_i be the event that vertex i is isolated in $G_{n,p}$. As vertex i is isolated if, and only if, $G_{n,p}$ contains none of the $n - 1$ edges $\{i, j\}$ with $j \neq i$, it follows that $P(A_i) = 2^{-(n-1)}$. More generally, any k vertices, say i_1, \dots, i_k , are isolated if, and only if, $G_{n,p}$ contains none of the $k(n-1) - \binom{k}{2}$ distinct possibilities for edges incident on them (each vertex has $n - 1$ possible edges incident upon it and the $\binom{k}{2}$ possible edges between the k vertices are double counted). Accordingly,

$$P(A_{i_1} \cap \dots \cap A_{i_k}) = (1-p)^{k(n-1)-\binom{k}{2}}.$$

The inclusion-exclusion summands $S_k = S_k^{(n)}$ hence satisfy

$$S_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \binom{n}{k} (1-p)^{k(n-1)-\binom{k}{2}}.$$

To determine the critical range of p we consider where S_k is essentially constant for large n . If $p = p_n$ tends to zero as $n \rightarrow \infty$ then the term $(1-p)^{-k-\binom{k}{2}}$ tends to one for each fixed k . In view of (6.4), we then obtain

$$S_k = (1 + \zeta_n) \frac{\{n(1-p)^n\}^k}{k!}$$

where $\zeta_n \rightarrow 0$ as $n \rightarrow \infty$. We are hence led to consider the behaviour of

$$n(1-p)^n = n \exp(n \log(1-p)) = n \exp(n(-p - p^2/2 - p^3/3 - \dots)).$$

The first term in the expansion of the logarithm dominates the rest if $p = p_n \rightarrow 0$ and we are hence led to select p so that ne^{-np} is approximately constant. Taking logarithms we quickly hone in on the critical rate of decay of p with n and set $p = p_n = \frac{\log n}{n} + \frac{c}{n}$ for a real constant c . Working backwards to verify each step in the preliminary analysis is now straightforward. We first check that $ne^{-np} = e^{-c}$ is constant. The remaining terms from the Taylor expansion of the logarithm contribute little as

$$\begin{aligned} \frac{p^2}{2} + \frac{p^3}{3} + \frac{p^4}{4} + \dots &= \frac{p^2}{2} \left(1 + \frac{2p}{3} + \frac{2p^2}{4} + \dots\right) \\ &< \frac{p^2}{2}(1 + p + p^2 + \dots) = \frac{p^2}{2(1-p)} \leq p^2 \end{aligned}$$

for $0 \leq p \leq 1/2$, the inequality hence certainly holding for all sufficiently large n when $p = \frac{\log n}{n} + \frac{c}{n}$. In consequence, for any fixed c , we eventually have

$$0 \leq n \left(\frac{p^2}{2} + \frac{p^3}{3} + \frac{p^4}{4} + \dots \right) < np^2 = \frac{(\log(n) + c)^2}{n} \rightarrow 0,$$

whence $e^{-n(p^2/2+p^3/3+\dots)} \rightarrow 1$ as $n \rightarrow \infty$. It follows that $S_k = S_k^{(n)} \rightarrow (e^{-c})^k/k!$ and the Poisson paradigm comes into play again.

THEOREM Suppose $p = p_n = \frac{\log n}{n} + \frac{c}{n}$ where c is an arbitrary real constant. Then the number of isolated vertices of the random graph $G_{n,p}$ is governed asymptotically by a Poisson distribution with parameter e^{-c} . A fortiori, the probability that there exist isolated vertices in $G_{n,p}$ tends to $1 - e^{-e^{-c}}$ as $n \rightarrow \infty$.

The limiting double exponential $e^{-e^{-c}}$ is a little startling to be sure but is otherwise unexceptionable—it is a constant and that is all that matters.

A very sharp threshold at a fine order of infinity is in evidence here: small changes in the constant c squirrelled away inside the subdominant term result in dramatic changes in isolation probabilities. If c is even slightly negative then the probability that there are isolated vertices is very close to one; if, on the other hand, c is just slightly positive then the probability that there are isolated vertices is very small. The transition from probabilities near one to probabilities near zero happens spectacularly fast: if $c = -5$ the probability that there are isolated vertices is in excess of $1 - 10^{-64}$; if $c = +5$ the probability that there are isolated vertices is less than 0.007. As c/n is subdominant compared to $\log(n)/n$, this says that *isolated vertices abound when p lies below $\log(n)/n$ and, lemming-like, suffer mass extinction when p increases just past $\log(n)/n$* .

8 Connectivity, from feudal states to empire

It is clear that a graph containing isolated vertices is not connected but the converse is not, in general, true. Nevertheless, Erdős and Rényi showed in a *tour de force* of 1960⁴ that the dominant mechanism due to which a random graph becomes disconnected asymptotically is because of the presence of isolated vertices.

A *component* of a graph is a connected subgraph on a maximal subcollection of vertices. The *order* of a component is the number of vertices it contains. Thus, an isolated vertex is a component of order 1, an isolated pair of vertices connected by an edge forms a component of order 2, and a connected graph is a component of order n .

Let $P_n^{(k)}$ denote the probability that the random graph $G_{n,p}$ contains a component of order k . The theorem of the previous section shows that when $p = p_n = \frac{\log n}{n} + \frac{c}{n}$ then $P_n^{(1)} \rightarrow 1 - e^{-e^{-c}}$ as $n \rightarrow 0$. What can be said about components of other orders?

⁴P. Erdős and A. Rényi, *op. cit.*

The probability that two given vertices form a component is $p(1-p)^{2(n-2)}$ as an edge must exist between the two vertices and each of these vertices must be isolated from the others. As there are $\binom{n}{2}$ ways of selecting the vertex pair, by Boole's inequality,

$$\begin{aligned} P_n^{(2)} &\leq \binom{n}{2} p(1-p)^{2(n-2)} < \frac{n^2}{2} \left(\frac{\log n + c}{n} \right) \left(1 - \frac{\log n + c}{n} \right)^{2n-4} \\ &\leq \frac{n \log n}{2} \left(1 + \frac{c}{\log n} \right) \exp \left(-2 \log n - 2c + \frac{4(\log n + c)}{n} \right) < \frac{\log n}{n} e^{-2c} \end{aligned}$$

eventually. The penultimate step follows by the elementary inequality $1+x \leq e^x$ valid for all real x ; the final step because the term $(1+c/\log n) \exp(4(\log n/n)(1+c/\log n))$ tends to 1 as $n \rightarrow \infty$ and, in particular, is less than 2 for all sufficiently large n . The specific constants in the upper bound are irrelevant—all that matters is that it vanishes asymptotically. It follows that $P_n^{(2)} \rightarrow 0$ as $n \rightarrow \infty$ and components of order 2 are rare.

A component of order 2 is simple as there is only one way in which two vertices can be connected by an edge. The picture is a little more nuanced for orders $3 \leq k \leq \lfloor n/2 \rfloor$. A given subcollection of $k \geq 3$ vertices forms a component if, and only if, each of these vertices is isolated from the remaining $n-k$ vertices and the subgraph of $G_{n,p}$ on these vertices is connected. As there are $k(n-k)$ possible edges between the given subcollection of k vertices and the remaining $n-k$ vertices, the probability that the given subcollection is isolated from the rest of the graph is just $(1-p)^{k(n-k)}$. To form a component of order k on these vertices it now only remains to ensure that the subgraph on these k vertices is connected. While the structure of connected graphs is complicated, we can finesse detailed calculations by realising that, as outlined in our discussion on trees in Section 3, each connected graph has at least one imbedded tree. Thus, we may associate with each connected graph an imbedded tree, it makes little matter which one. Each tree T' on the k vertices may perforce be associated with an equivalence class $\{\mathcal{G}'\}$ of connected subgraphs on these vertices, each of which contains the tree in question. The occurrence of any of the subgraphs \mathcal{G}' in this equivalence class certainly implies the occurrence of the associated imbedded tree T' and it follows that the probability that $G_{n,p}$ contains any of these connected subgraphs is bounded above by the probability that $G_{n,p}$ contains the tree T' . As each tree on k vertices contains exactly $k-1$ edges (Theorem 3.2), the probability that a given tree T' is generated is exactly p^{k-1} and this then bounds from above the probability that $G_{n,p}$ contains any of the connected subgraphs \mathcal{G}' associated with this tree. As we allow T' to vary over all trees on the k vertices, we end up accounting for all connected subgraphs on these vertices. As, by Cayley's formula, there are exactly k^{k-2} trees on k vertices, the probability that a component is formed on a given set of k vertices is hence bounded above by $k^{k-2} p^{k-1} (1-p)^{k(n-k)}$. By considering all possible subcollections of k vertices, we obtain $P_n^{(k)} \leq \binom{n}{k} k^{k-2} p^{k-1} (1-p)^{k(n-k)}$ by another application of Boole's inequality.

The expression is certainly messier than was the case of order 2 but if the reader keeps her eye on the critical terms it will resolve readily enough. In view of (5.1) and (5.2), we may write $\binom{n}{k} < n^k k^{-k} e^k$ so that

$$\begin{aligned} \binom{n}{k} k^{k-2} p^{k-1} &< nk^{-2} (np)^{k-1} e^k = nk^{-2} \exp \{ (k-1) \log(\log n + c) + k \} \\ &< nk^{-2} \exp \{ k(\log(\log n + c) + 1) \} \end{aligned}$$

eventually as, even if the constant c is negative, $\log n + c$ becomes larger than 1 for all sufficiently large n . Now for $k \leq n/2$ we may make the elementary observation that $(n - k)p \geq np/2 = (\log n + c)/2$. Again, in view of the inequality $1 + x \leq e^x$, we may hence bound the remaining term by

$$(1 - p)^{k(n-k)} \leq \exp(k(n-k)(-p)) \leq \exp(-k(\log n + c)/2).$$

Pooling expressions we obtain $P_n^{(k)} < nk^{-2} \exp\{-k(\frac{1}{2}\log n - \zeta_n)\}$ where we lump the subdominant expressions into the term $\zeta_n = \log(\log n + c) - c/2 + 1 = \log \log n + \log(1 + c/\log n) - c/2 + 1$. In view of the slow growth of $\log \log n$ compared to $\log n$, it is clear that ζ_n is dominated by $\log n$ asymptotically and, in particular, $\zeta_n < \log(n)/8$ for all sufficiently large n .

The reader may be forgiven for being bemused by the selection of the apparently arbitrary constant $1/8$ as it seems that one could (for large enough n) have set $\zeta_n < r \log n$ for any strictly positive r less than one. The necessity for some care in the selection of the constant becomes apparent if she keeps an eye on the bigger picture. If the bound on $P_n^{(k)}$ is to be non-trivial then the large multiplying factor n must be dominated by the exponential term. But this then means that $k(\frac{1}{2} - r)$ must be strictly larger than 1. As $k \geq 3$ this means that, in the worst case, we require $3(\frac{1}{2} - r) > 1$ or $r < 1/6$. The choice $r = 1/8$ keeps the fractions simple. Magic is always trite once the sleight of hand is explained.

So we now have $P_n^{(k)} < nk^{-2} \exp\{-3(\frac{1}{2} - \frac{1}{8})\log n\} = n^{-1/8}k^{-2}$ for all $k \geq 3$. By summing over all orders from 3 to $\lfloor n/2 \rfloor$, the probability that $G_{n,p}$ contains one or more components with orders between 3 and $\lfloor n/2 \rfloor$ may now be overbounded via Boole's inequality by

$$\sum_{k=3}^{\lfloor n/2 \rfloor} P_n^{(k)} < n^{-1/8} \sum_{k=3}^{\lfloor n/2 \rfloor} k^{-2} \leq n^{-1/8} \sum_{k=1}^{\infty} k^{-2}.$$

The series on the right is convergent, for example, by the integral test, as

$$\sum_{k=1}^{\infty} k^{-2} = 1 + \sum_{k=2}^{\infty} k^{-2} < 1 + \int_1^{\infty} x^{-2} dx = 2.$$

It follows that $\sum_{k=3}^{\lfloor n/2 \rfloor} P_n^{(k)} < 2n^{-1/8} \rightarrow 0$ as $n \rightarrow \infty$.

If the random graph $G_{n,p}$ contains no components of orders between 2 and $\lfloor n/2 \rfloor$ then it must be the case that it consists of one giant component of order $k > n/2$ together with $n - k$ isolated vertices. If, in addition, it contains no isolated vertices then it must follow that the single giant component that remains must have order n , that is to say, the graph is connected. It follows that

$$P\{G_{n,p} \text{ is connected}\} = 1 - \sum_{k=1}^{\lfloor n/2 \rfloor} P_n^{(k)}.$$

But, as we have seen, $P_n^{(1)} + P_n^{(2)} + \sum_{k=3}^{\lfloor n/2 \rfloor} P_n^{(k)} \rightarrow 1 - e^{-e^{-c}}$ as $n \rightarrow \infty$. Following in the path of Erdős and Rényi, we have discovered a remarkable theorem.

⁵The reader familiar with Euler's solution of Pietro Mengoli's *Basel Problem* will know indeed that $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$ but we won't stop to pull this beguiling chestnut out of the fire.

THEOREM 1 If $p = p_n = \frac{\log n}{n} + \frac{c}{n}$ for some constant c , then the probability that the random graph $G_{n,p}$ is connected tends to $e^{-e^{-c}}$ as $n \rightarrow \infty$.

If c is strictly negative then the probability that the graph is connected is very small; if, on the other hand, c is strictly positive then the probability that the graph is connected is well-nigh certain.

Now it is intuitive that connectivity is a *monotone property* in the sense that the chance that $G_{n,p}$ is connected improves as p increases.

THEOREM 2 If $p_1 < p_2$ then $\mathbf{P}\{G_{n,p_1} \text{ is connected}\} \leq \mathbf{P}\{G_{n,p_2} \text{ is connected}\}$.

PROOF: We begin with the trite observation that if \mathcal{G} and \mathcal{G}' are graphs on n vertices with the edges of \mathcal{G} contained among the edges of \mathcal{G}' then, if \mathcal{G} is connected, so is \mathcal{G}' . Set $p = (p_2 - p_1)/(1 - p_1)$ and consider independently generated random graphs $G_1 = G_{n,p_1}$ and $G = G_{n,p}$. By merging G_1 and G_2 we now form a larger new graph G_2 on n vertices, each edge of G_2 being either an edge of G_1 or an edge of G or both. As G_2 is larger than G_1 , it is clear that $\mathbf{P}\{G_1 \text{ is connected}\} \leq \mathbf{P}\{G_2 \text{ is connected}\}$.

Now, by the simplest version of inclusion and exclusion, the probability that a given edge is in G_2 is given by $p_1 + p - p_1 p = p_2$. As by independent generation of the edges of G_1 and G_2 it is clear that the edges of G_2 are also generated independently, it follows that G_2 is an instance of the random graph G_{n,p_2} . ▶

In view of the monotone nature of connectivity, we may present Theorem 1 in a more vivid language: if $p_n \ll \log(n)/n$ then, for sufficiently large n , $p_n \leq (\log n - 10^{100})/n$ and G_{n,p_n} is almost certainly not connected; if, on the other hand, $p_n \gg \log(n)/n$ then, for sufficiently large n , $p_n \geq (\log n + 10^{100})/n$ and G_{n,p_n} will be almost certainly connected.

THEOREM 3 Suppose $\{p_n\}$ is a sequence of probabilities. Then, as $n \rightarrow \infty$,

$$\mathbf{P}\{G_{n,p_n} \text{ is connected}\} \rightarrow \begin{cases} 0 & \text{if } \frac{p_n}{\log(n)/n} \rightarrow 0, \\ 1 & \text{if } \frac{p_n}{\log(n)/n} \rightarrow \infty. \end{cases}$$

Thus connectivity emerges abruptly when $p = p_n$ is of the order of $\log(n)/n$. The random graph theorist says that $\log(n)/n$ is a *threshold function* (the physicist would call it a *phase transition*) for the property of graph connectivity.

A dynamical model of graph generation emphasises this sudden emergence. The random graph $G_{n,p}$ contains close to $p \binom{n}{2}$ edges with high probability. (The nature of the approximation is made precise by the law of large numbers of Section V.6 and Problem V.17, but an intuitive feel is sufficient for our purposes here.) Now consider the choice $p = p_n = (\log n + c)/n$. If $c = 0$ there are close to $\frac{1}{2}n \log n$ edges; if $c < 0$ there are much fewer than $\frac{1}{2}n \log n$ edges; and if $c > 0$ there are much more than $\frac{1}{2}n \log n$ edges. Now consider a dynamical graph process \mathcal{G}_t which starts at time zero with the graph \mathcal{G}_0 on n vertices with no edges and, at each instant, adds a randomly chosen edge from those not yet selected. Our theorem then says that when $t \ll \frac{1}{2}n \log n$ the graph is highly fragmented and consists of many isolated vertices but when $t \gg \frac{1}{2}n \log n$ all the

isolated components have merged into a connected piece. As the graph evolves from just before to just past the critical time $\frac{1}{2}n \log n$ it moves, in the blink of an eye in this time frame, from a highly feudal system of independent city states to an empire of connected elements. All roads do indeed lead to Rome. Eventually.

9 Sieves, the Lovász local lemma

We consider a general setting where A_1, \dots, A_n are events in an *arbitrary* probability space. If we describe the events A_1, \dots, A_n picturesquely, if perhaps slightly inaccurately, as “bad” events, in many applications we are interested in the occurrence of the “good” event $A_1^c \cap \dots \cap A_n^c$. The reader has seen this theme repeatedly in the previous sections. But in such settings it may not be immediately apparent whether the “good” event can occur—if the “bad” events cover the space then it is certain that one or more of them will occur and if the dependency structure among the A_i is complex then it may not at all be trivial to determine whether this is the case. The situation is akin to that of a putative needle in a probabilistic haystack with the “good” event identified with the needle. A probability sieve is a metaphorical construct which attempts to isolate the problematic needle: if we can construct an experiment in which it can be demonstrated that the “good” event occurs with strictly positive probability then we can deduce the existence of the needle.

The simplest sieve undoubtedly arises directly from Boole’s inequality. We have $P(\bigcap_i A_i^c) = 1 - P(\bigcup_i A_i) \geq 1 - \sum_i P(A_i)$, the inequality leading to a simple conclusion.

BOOLE’S SIEVE *If $\sum_i P(A_i) < 1$ then $P(\bigcap_i A_i^c) > 0$.*

As we saw in the application of Boole’s inequality to Ramsey theory, the strength of Boole’s sieve lies in the fact that it makes no assumptions about the relationships between the A_i . On the other hand, the tool is rather crude and the probabilities of the A_i have to be quite small for it to be effective. We can move away from the constraint of small *individual* event probabilities by imposing a structural independence constraint on the *family* of events as we see in the following theorem.

INDEPENDENCE SIEVE *If A_1, \dots, A_n are independent and $P(A_i) < 1$ for each i then $P(\bigcap_i A_i^c) > 0$.*

PROOF: The events A_1^c, \dots, A_n^c inherit independence from the independence of A_1, \dots, A_n and so $P(\bigcap_i A_i^c) = \prod_i (1 - P(A_i)) > 0$ as each of the terms in the product is strictly positive. ▶

An intermediate situation between Boole’s sieve and the independence sieve arises if there is *some* dependency structure in the A_i as seen, for example,

in the applications of the Poisson paradigm in the previous section. These applications were all characterised by an “asymptotic independence”, however, so that the situations were near the independence sieve end of the scale. A new principle is going to be needed if the dependences are more severe. Some notation first.

A dependency graph \mathcal{G} of the events A_1, \dots, A_n is a graph on n vertices (corresponding to the indices of the A_i). The edges of \mathcal{G} are determined by the dependency structure of the A_i as follows. For each i , let the *independence set* J_i be the maximal set of indices j for which A_i is independent of the set of events $\{A_j : j \in J_i\}$. In other words, J_i is the maximal set with the property $P(A_i \cap \bigcap_{j \in J} A_j) = P(A_i) P(\bigcap_{j \in J} A_j)$ for every subset J of J_i . Then $\{i, j\}$ is an edge of the dependency graph \mathcal{G} if, and only if, j is *not* in J_i . For each i , the vertices of the graphs are hence partitioned into two disjoint groups as sketched in Figure 4 which shows that piece of a dependency graph consisting of the edges emanating from vertex i . The reader should keep Bernstein’s construction in

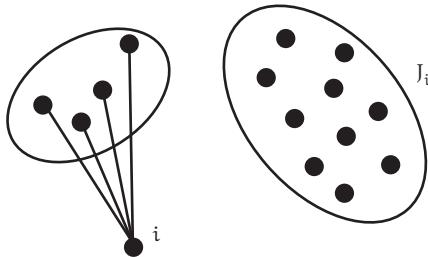


Figure 4: The independence set of a given vertex.

Example III.2.7 in mind: the joint independency property defining the index sets J_i is much more stringent than merely pairwise independence. Thus, the pairwise independence of A_i and A_j is not, in itself, sufficient to place j in J_i .

For each vertex i , the *vertex degree* $\deg(i)$ is the number of edges of the dependency graph that are incident on i . We will suppose initially that $P(A_i) \leq p$ and $\deg(i) \leq d$ for each i .

To turn to the event of interest, the chain rule for conditional probabilities allows us to write

$$P(A_1^c \cap \dots \cap A_n^c) = \prod_{i=1}^n P(A_i^c \mid A_{i-1}^c \cap \dots \cap A_1^c) = \prod_{i=1}^n [1 - P(A_i \mid A_{i-1}^c \cap \dots \cap A_1^c)].$$

Now, $P(\bigcap_i A_i^c) > 0$ if, and only if, $P(A_i \mid A_{i-1}^c, \dots, A_1^c)$ is strictly less than 1 for each i . This suggests that we attempt to bound conditional probabilities of the form $P(A_i \mid \bigcap_{j \in V} A_j^c)$ where V denotes any index set of vertices. It clearly suffices if we can find some positive α strictly less than one for which

$$P(A_i \mid \bigcap_{j \in V} A_j^c) \leq \alpha$$

for each i and every subset of vertices V . This looks formidable. Lovász's clever idea out of the impasse was to attempt an induction on the size of V to reduce the problem to bite-sized pieces.

The induction base is trivial. When V is empty we have

$$\mathbf{P}(A_i \mid \bigcap_{j \in V} A_j^c) = \mathbf{P}(A_i \mid \Omega) = \mathbf{P}(A_i) \leq p.$$

(The reader should recall the usual convention that an intersection over the empty set is the entire space.) The induction base is hence established for any choice $\alpha \geq p$.

Now suppose V is any subset of v vertices; we may suppose $i \notin V$ as else the conditional probability is trivially zero. Let $V'_i := V \cap J_i^c$ be the vertices in V that are adjacent to i , i.e., V'_i consists of those vertices j in V for which $\{i, j\}$ is an edge of the dependency graph, and let $V_i := V \cap J_i = V \setminus V'_i$ be the remaining vertices in V . We observe that A_i is independent of the collection of events $\{A_j^c, j \in V_i\}$ as V_i is a subset of the independence set J_i . The hypothesis is immediate if $V'_i = \emptyset$ (i.e., $V_i = V$) so let us focus on the case when V_i contains $v - 1$ or fewer vertices. We can separate the sets V'_i and V_i by a simple conditioning argument:

$$\mathbf{P}(A_i \mid \bigcap_{j \in V} A_j^c) = \frac{\mathbf{P}(A_i \cap \bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c)}{\mathbf{P}(\bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c)}.$$

We bound the numerator by first eliminating the events A_j^c ($j \in V'_i$) by exploiting monotonicity of probability measure and then using the fact that A_i is independent of the events A_k^c ($k \in V_i$) to obtain

$$\mathbf{P}(A_i \cap \bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c) \leq \mathbf{P}(A_i \mid \bigcap_{k \in V_i} A_k^c) = \mathbf{P}(A_i) \leq p.$$

And as for the denominator,

$$\begin{aligned} \mathbf{P}(\bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c) &= 1 - \mathbf{P}(\bigcup_{j \in V'_i} A_j \mid \bigcap_{k \in V_i} A_k^c) \\ &\geq 1 - \sum_{j \in V'_i} \mathbf{P}(A_j \mid \bigcap_{k \in V_i} A_k^c) \geq 1 - \alpha d, \end{aligned}$$

the last step following via the induction hypothesis (as V_i contains fewer than v vertices) and the assumed fact $\deg(i) \leq d$. It follows that

$$\mathbf{P}(A_i \mid \bigcap_{j \in V} A_j^c) \leq \frac{p}{1 - \alpha d}$$

and the bound on the right is no larger than α for values of α for which the quadratic $Q(\alpha) = \alpha^2 d - \alpha + p$ is negative. If $4pd \leq 1$ then the discriminant

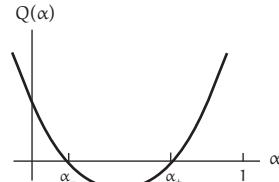


Figure 5: When a quadratic has real roots.

$\Delta = 1 - 4pd$ of the quadratic is positive and $Q(\alpha)$ will have two real roots, $\alpha_{\pm} = [1 \pm \sqrt{1 - 4pd}] / 2d$ (Figure 5), which are positive and less than 1. (The roots coincide if $4pd = 1$.)

We may select any α in the range $0 < \alpha_- \leq \alpha \leq \alpha_+ < 1$ to complete the induction. While α may be arbitrarily selected in this range, we may for definiteness select $\alpha = 2p$. (That this choice of α falls in the right range may be seen from the simple observation $1 - \sqrt{1 - x} < x < 1 + \sqrt{1 - x}$ valid for all $0 < x < 1$.) Thus, if $4pd \leq 1$ then $P(A_i | \bigcap_{j \in V} A_j^c) \leq 2p$ for all subsets of indices V and we have proved the deceptively simple

LOVÁSZ LOCAL LEMMA *If $4pd \leq 1$ then $P(A_1^c \cap \dots \cap A_n^c) > 0$.*

It is useful to think of the local lemma as filling in a continuum of sieves between the extremes of Boole's sieve, which requires small individual event probabilities but imposes no constraint on the family of events, and the independence sieve, which allows large individual event probabilities but imposes a strong independence constraint on the family of events. The local lemma allows a trade-off between the event probabilities and the dependency structure.

The local lemma was published by Laslo Lovász and Paul Erdős in 1975⁶ and, elementary proof notwithstanding, it is impossible to overestimate the abiding impact it has had.



REFINEMENTS

A close examination of the proof of the local lemma shows that mutual independence between events not connected by edges in the dependency graph is not, strictly speaking, essential for the result to hold. To generalise, suppose $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a graph on vertices $\mathcal{V} = \{1, \dots, n\}$ corresponding to the indices of the events A_1, \dots, A_n . The edge-set \mathcal{E} connects vertices corresponding to dependent events as before with the proviso that vertices not connected by edges are not necessarily independent but may be weakly dependent in the sense below. For each i , denote by J_i the set of vertices j that are non-adjacent to i , that is to say, $J_i = \{j : \{i, j\} \notin \mathcal{E}\}$.

LOVÁSZ LOCAL LEMMA (INDUSTRIAL STRENGTH) *Suppose x_1, \dots, x_n are positive numbers all strictly less than 1. If*

$$P(A_i | \bigcap_{j \in J} A_j^c) \leq x_i \prod_{j \notin J_i} (1 - x_j)$$

for each i and every subset J of J_i , then

$$P(A_1^c \cap \dots \cap A_n^c) \geq \prod_{i=1}^n (1 - x_i),$$

and, in particular, the event that none of the A_i occur has strictly positive probability.

⁶P. Erdős and L. Lovász, "Problems and results on 3-chromatic hypergraphs and some related questions", in A. Hajnal et al. (eds), *Infinite and Finite Sets*, pp. 609–628. Amsterdam: North-Holland, 1975.

PROOF: It will suffice to show that $P(A_i \mid \bigcap_{j \in V} A_j^c) \leq x_i$ for each i and every subset of vertices V . We again proceed by induction on the size of V .

Let V be any subset of v vertices; we may again assume that $i \notin V$ as the conditional probability is trivially zero in this case. Write $V'_i = V_i \cap J_i^c$ for the vertices in V that are adjacent to i , i.e., V_i consists of those vertices j in V for which $\{i, j\}$ is an edge of the graph G , and let $V_i := V \setminus J_i = V \setminus V_i$ be the remaining vertices in V . The hypothesis is immediate if $V'_i = \emptyset$ (i.e., $V_i = V$) so let us focus on the case when V_i contains $v - 1$ or fewer vertices. As before, we separate the sets V_i and V'_i by conditioning to obtain

$$P(A_i \mid \bigcap_{j \in V} A_j^c) = \frac{P(A_i \cap \bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c)}{P(\bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c)}.$$

As before, we bound the numerator by exploiting monotonicity of probability measure to eliminate the events A_j^c ($j \in V'_i$) and obtain

$$P\left(A_i \cap \bigcap_{j \in V'_i} A_j^c \mid \bigcap_{k \in V_i} A_k^c\right) \leq P\left(A_i \mid \bigcap_{k \in V_i} A_k^c\right) \leq x_i \prod_{j \notin J_i} (1 - x_j)$$

by the stated conditions of the lemma. We will need to exercise a little more care with the denominator. Suppose $V = \{j_1, \dots, j_v\}$ with the elements ordered for simplicity so that $V'_i = \{j_1, \dots, j_r\}$ and $V_i = \{j_{r+1}, \dots, j_v\}$. Factoring via the chain rule we then obtain

$$\begin{aligned} P\left(\bigcap_{s=1}^r A_{j_s}^c \mid \bigcap_{t=r+1}^v A_{j_t}^c\right) &= \prod_{s=1}^r P\left(A_{j_s}^c \mid \bigcap_{t=s+1}^v A_{j_t}^c\right) \\ &= \prod_{s=1}^r \left[1 - P\left(A_{j_s} \mid \bigcap_{t=s+1}^v A_{j_t}^c\right)\right] \geq \prod_{s=1}^r (1 - x_{j_s}) \geq \prod_{j \notin J_i} (1 - x_j), \end{aligned}$$

the penultimate step following via repeated application of the induction hypothesis to the multiplicands. The induction is complete. ▶

A consideration of the “baby” version of the local lemma will help demystify the x_i ’s in the lemma which at first blush appear somewhat mysterious. We recall first that the Taylor series for the logarithm $\log(1 - z) = -z - \frac{1}{2}z^2 - \frac{1}{3}z^3 - \dots$ is convergent for all $|z| < 1$. It follows that

$$\begin{aligned} \left(1 - \frac{1}{n}\right)^{n-1} &= \exp\left\{(n-1) \log\left(1 - \frac{1}{n}\right)\right\} = \exp\left\{-\frac{n-1}{n} - \frac{n-1}{2n^2} - \dots\right\} \\ &= \exp\left\{-1 + \frac{1}{n} - \frac{n-1}{2n^2} - \frac{n-1}{3n^3} - \dots\right\} \end{aligned}$$

for all $n > 1$. Expressing the right-hand side in the form $e^{-1+\zeta}$ we see that

$$\begin{aligned} \zeta &= \frac{1}{n} - (n-1) \sum_{k=2}^{\infty} \frac{1}{kn^k} = \frac{1}{n} - \frac{n-1}{2n^2} \sum_{k=0}^{\infty} \frac{2}{(k+2)n^k} \\ &> \frac{1}{n} - \frac{n-1}{2n^2} \sum_{k=0}^{\infty} \frac{1}{n^k} = \frac{1}{n} - \frac{n-1}{2n^2} / \left(1 - \frac{1}{n}\right) = \frac{1}{n} - \frac{1}{2n} = \frac{1}{2n} \end{aligned}$$

is positive. In consequence we have the elementary inequality $(1 - 1/n)^{n-1} > 1/e$ valid whenever $n > 1$.

To return to the local lemma, suppose that the maximum degree of the graph is bounded by d , that is to say, for each i , $\deg(i) \leq d$. Furthermore, for each i and subset J of J_i , suppose $P(A_i | \bigcap_{j \in J} A_j^c) \leq p$. We observe that in the case when A_i is independent of the set $\{A_j, j \in J\}$ this reduces to the condition of the “baby” version of the lemma, to wit $P(A_i) \leq p$ for each i . If we set $x_i = 1/(d+1)$ for each i then

$$x_i \prod_{j \notin J_i} (1 - x_j) \geq \frac{1}{d+1} \left(1 - \frac{1}{d+1}\right)^d > \frac{1}{e(d+1)}.$$

It follows that *the conclusion of the local lemma remains valid if $ep(d+1) < 1$* . The reader should note in passing that we have improved the constant 4 in the “baby” version of the lemma to e .

The reader who is interested will find a more subtle sieve developed in Chapter XVIII; see also Problems XVIII.20–23.

10 Return to Ramsey theory

Let us put the local lemma to work in Ramsey theory in the context of the van der Waerden numbers defined in Section 5. Boole’s inequality provided the lower bound $W(k) > 2^{k/2}$. Can the local lemma improve this estimate?

As before, we randomly two-colour the integers from 1 to n . Let S be any embedded k -term arithmetic sequence, A_S the event that S is monochromatic. It is clear that $P(A_S) = 2^{1-k}$ does not depend on the particular choice of S . Let us now consider the dependency graph G induced by the events $\{A_S\}$. The vertices of G are indexed by the various k -term arithmetic sequences S that are embedded in the integers from 1 to n . Two events A_S and A_T are independent if, and only if, the arithmetic sequences S and T do not share any elements; more generally, the event A_S is independent of the family of events $\{A_T, T \in \mathcal{T}\}$ if, and only if, S shares no elements with any arithmetic sequence T in the family. So we come to a basic question: how many arithmetic sequences intersect S ?

Let us begin by estimating the number of sequences that intersect a given point x of S . Call the distance between successive elements of an arithmetic sequence its *span*. Now suppose a k -term arithmetic sequence T of span g intersects S at x . By shifting T left or right g units at a time we can systematically enumerate all possible k -term sequences of span g that intersect S at x (though not all of them may be confined to the integers from 1 to n and some of them may intersect other members of S as well). It follows that there are at most k arithmetic sequences of length k and span g that intersect S at x (see Figure 6). The largest span of any k -term arithmetic sequence in $\{1, \dots, n\}$ is $\lfloor n/k \rfloor$ so that $1 \leq g \leq n/k$. It follows that there are at most $k \times n/k = n$ arithmetic sequences of length k that intersect S at x . As there are k possible values for x we conclude

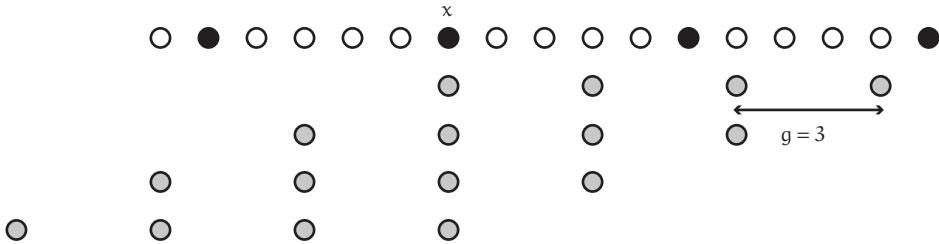


Figure 6: Arithmetic sequences containing a given point.

that there are at most nk arithmetic sequences T that intersect S . The vertex degree of the dependency graph is hence $\leq nk$.

All is set for the Lovász local lemma with $p = 2^{1-k}$ and $d = nk$. If $n \leq 2^k/(8k)$ then, with strictly positive probability, a random colouring of the numbers from 1 to n will contain no monochromatic k -term arithmetic sequence; this implies the existence of at least one such colouring. We have isolated a kernel of wheat from an intolerable sack of meal.

THEOREM 1 *For each $k \geq 1$, we have $W(k) > 2^k/(8k)$.*

The improvement in the estimate $2^{k/2}$ of Theorem 5.3 may be traced directly to the fact that the degree $\leq nk$ of the dependency graph is modest, relatively speaking, compared to the number of k -term arithmetic sequences embedded in 1 to n . As we have already seen in deploying Boole's sieve, the number of arithmetic sequences is of the order of n^2 where we recall that n grows at least exponentially fast with k .

The improvement in the lower bound for the van der Waerden numbers is impressive but honesty compels me to mention that, using purely non-probabilistic methods, Elwyn Berlekamp showed in 1968 that $W(p+1) > p2^p$ for any prime p . Nonetheless, the reader should remark on the almost effortless deployment of the local lemma and the power and elegance of the results that follow. In many instances the results given by the local lemma are not merely good but are not matched by any other method. When applied to off-diagonal Ramsey numbers, for instance, the local lemma yields the best lower bounds known (*circa* 2012).



11 Latin transversals and a conjecture of Euler

Let $A = [a_{ij}]$ be a square matrix of order n . A permutation π of the integers from 1 to n is called a *Latin transversal* (of A) if the entries $a_{i\pi(i)}$ ($1 \leq i \leq n$) are all distinct. Thus, if

$$A = \begin{bmatrix} x & x & z \\ y & z & z \\ x & x & x \end{bmatrix},$$

then the permutation $\pi(1) = 3, \pi(2) = 1, \pi(3) = 2$ is a Latin transversal as

$$(a_{1\pi(1)}, a_{2\pi(2)}, a_{3\pi(3)}) = (z, y, x)$$

is comprised of distinct elements of A . Intuitively, if no element in A appears too frequently then we should be able to find a Latin transversal. Can we codify this?

Suppose no element appears m or more times in A . Let \mathcal{V} denote the family of quadruples of indices (i, j, i', j') with $i < i'$ and $j \neq j'$ for which $a_{ij} = a_{i'j'}$. Each quadruple (i, j, i', j') in \mathcal{V} identifies a distinct pair of cell locations (i, j) and (i', j') in the array A , at which locations the entries are identical. A permutation π will not lead to a Latin transversal if $\pi(i) = j$ and $\pi(i') = j'$ for some (i, j, i', j') in \mathcal{V} .

Auxiliary randomisation is again indicated in our search for a good permutation. Suppose π is a *random permutation* chosen uniformly from the $n!$ possible permutations of the integers from 1 to n . Let $A_{iji'j'}$ denote the subset of permutations π satisfying $\pi(i) = j$ and $\pi(i') = j'$. If (i, j, i', j') is in \mathcal{V} then $A_{iji'j'}$ connotes a “bad” event. In this terminology, if the “good” event $\bigcap_{\mathcal{V}} A_{iji'j'}^c$ occurs, then a Latin transversal has been found. It follows that a Latin transversal exists if $P(\bigcap_{\mathcal{V}} A_{iji'j'}^c) > 0$. The stage is set for the Lovász local lemma.

We consider a graph \mathcal{G} with vertices indexed by the quadruples (i, j, i', j') in \mathcal{V} . Two quadruples (i, j, i', j') and (p, q, p', q') in \mathcal{V} are adjacent (i.e., are connected by an edge of the graph) if $\{i, i'\} \cap \{p, p'\} \neq \emptyset$ or $\{j, j'\} \cap \{q, q'\} \neq \emptyset$. The negation of the graph adjacency condition can be recast slightly more compactly: (i, j, i', j') and (p, q, p', q') are *not* adjacent if the four cells (i, j) , (i', j') , (p, q) , and (p', q') occupy four distinct rows and four distinct columns of A . The maximum degree of the graph is modest if no element appears too frequently in A . Indeed, for each quadruple (i, j, i', j') , there are no more than $4n$ choices of (p, q) with $p \in \{i, i'\}$ or $q \in \{j, j'\}$, and for each such choice of cell (p, q) there are fewer than m cells (p', q') with $a_{pq} = a_{p'q'}$. Excluding the quadruple (i, j, i', j') itself from the count, the maximum degree of \mathcal{G} is hence $\leq d = 4nm - 1$.

For any quadruple (i, j, i', j') with $i < i'$ and $j \neq j'$,

$$P(A_{iji'j'}) = \frac{(n-2)!}{n!} = \frac{1}{n(n-1)}$$

as there are $(n-2)!$ permutations with $\pi(i) = j$ and $\pi(i') = j'$. What can be said about the dependency structure of these events? While there is some dependence between events even when the corresponding vertices are not adjacent (because the permutation π corresponds to sampling without replacement), nonetheless the dependency is fairly mild. We claim indeed that

$$P\left(A_{iji'j'} \mid \bigcap_{\mathcal{U}} A_{pq,p'q'}^c\right) \leq \frac{1}{n(n-1)} \quad (11.1)$$

for any quadruple (i, j, i', j') in \mathcal{V} and every subset \mathcal{U} of quadruples (p, q, p', q') in \mathcal{V} that are not adjacent to (i, j, i', j') . This is the key to the analysis but before launching into a proof of (11.1) it will be useful to introduce some notation to keep the marauding subscripts under control.

Write $H = \bigcap_{\mathcal{U}} A_{pq,p'q'}^c$ for the set of permutations π which satisfy $\pi(p) \neq q$ and $\pi(p') \neq q'$ for every quadruple (p, q, p', q') in \mathcal{U} . It should be borne in mind that

H depends on \mathcal{U} and hence, implicitly, on the indices i , i' , j , and j' [as the quadruples (p, q, p', q') in \mathcal{U} cannot be adjacent to (i, j, i', j')]. We will find it convenient to introduce a nonce terminology for the elements of H : we say that a permutation π is *admissible* if it is in H . Conditioned on H , each of the admissible permutations has the same probability of being picked as the others.

Suppose ℓ and ℓ' are any two distinct indices. We suppress i and i' temporarily in the notation and write simply $A_{\ell\ell'} = A_{i\ell i'\ell'}$ to focus on ℓ and ℓ' . The event $A_{\ell\ell'} \cap H$ now represents the set of admissible permutations π satisfying $\pi(i) = \ell$ and $\pi(i') = \ell'$. A little thought on the manner of specification of H may suggest to the reader that, conditioned on H , the particular set $A_{jj'}$ with $\ell = j$ and $\ell' = j'$ may take on a special character. And indeed it does. We claim that $\text{card } A_{jj'} \cap H \leq \text{card } A_{\ell\ell'} \cap H$ for any pair $\ell \neq \ell'$. To prove the claim it will suffice to exhibit a one-to-one map from $A_{jj'} \cap H$ into $A_{\ell\ell'} \cap H$ and a little consideration shows that a simple transposition does the job.

THE SWAP (FIGURE 7): Let π be any admissible permutation in $A_{jj'}$. If $\ell \neq j$ then there exists some $k \neq i$ with $\pi(k) = \ell$; likewise, if $\ell' \neq j'$ there exists some $k' \neq i'$ with $\pi(k') = \ell'$. By simply transposing the elements of π at locations i , i' , k , and k' and keeping the other elements intact we generate a new permutation π^* with $\pi^*(i) = \ell$, $\pi^*(k) = j$, $\pi^*(i') = \ell'$, $\pi^*(k') = j'$, and $\pi^*(r) = \pi(r)$ for all other indices r . By construction, π^* is a unique permutation in $A_{\ell\ell'}$ so that the map $\pi \mapsto \pi^*$ is one-to-one. We claim moreover that π^* is admissible. Indeed, suppose (p, q, p', q') is not adjacent to (i, j, i', j') . Then $p \neq i$, $q \neq j$, $p' \neq i'$, and $q' \neq j'$. If $p \neq k$ then, by construction, $\pi^*(p) = \pi(p) \neq q$ (as π is admissible). If $p = k$, then $\pi^*(k) = j \neq q$. Arguing in this fashion for p' , we see that $\pi^*(p') \neq q'$. Thus, $\pi^*(p) \neq q$ and $\pi^*(p') \neq q'$ for every quadruple (p, q, p', q') in \mathcal{U} , that is to say, π^* is admissible. Thus, to each admissible permutation π in $A_{jj'}$ we can associate a unique admissible permutation π^* in $A_{\ell\ell'}$ and in consequence $\text{card } A_{jj'} \cap H \leq \text{card } A_{\ell\ell'} \cap H$ for any pair $\ell \neq \ell'$. We recall that, conditioned on H , all admissible permutations have the same probability. It follows that, for every $\ell \neq \ell'$, the conditional probability of $A_{jj'}$ given H is less than or equal to the conditional probability of $A_{\ell\ell'}$ given H , or, in notation, $P(A_{jj'} | H) \leq P(A_{\ell\ell'} | H)$. Summing both sides over all $n(n-1)$ pairs $\ell \neq \ell'$ results in

$$n(n-1) P(A_{jj'} | H) \leq \sum_{\ell \neq \ell'} P(A_{\ell\ell'} | H).$$

Now the sets $\{A_{\ell\ell'}, \ell \neq \ell'\}$ partition H , i.e., they are mutually disjoint and satisfy $\bigcup_{\ell \neq \ell'} A_{\ell\ell'} = H$. It follows via total probability that

$$\sum_{\ell \neq \ell'} P(A_{\ell\ell'} | H) = 1.$$

Or, reintroducing the moribund indices i and i' , $n(n-1) P(A_{iji'j'} | H) \leq 1$, which is the result (11.1) to be established.

We're almost there. An elementary inequality helps smooth the path.

LEMMA Set $f(x) = e^x(1 - x/(x+1))$ for $x \geq 0$. Then $f(x) \geq 1$ for all $x \geq 0$.

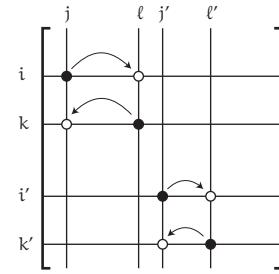


Figure 7: The swap.

PROOF: It is clear that $f(0) = 1$. An easy differentiation now shows that

$$f'(x) = e^x \left[1 - \frac{x}{x+1} - \frac{1}{x+1} + \frac{x}{(x+1)^2} \right] = e^x \frac{x}{(x+1)^2} \geq 0$$

for all $x \geq 0$. The claimed result follows quickly from the mean value theorem from elementary calculus which asserts that for any $x \geq 0$, $f(x) = f(0) + f'(z)x$ for some $0 \leq z \leq x$. \blacktriangleright

It follows as an easy consequence that $e^{-1/d} \leq 1 - 1/(d+1)$ for every positive integer d . Recalling now that the maximum degree of the graph \mathcal{G} is bounded above by $d = 4nm - 1$, if $m \leq (n-1)/4e$ then

$$\frac{1}{n(n-1)} = \frac{4m}{(d+1)(n-1)} \leq \frac{1}{e(d+1)} \leq \frac{1}{d+1} \left(1 - \frac{1}{d+1}\right)^d$$

and the Lovász local lemma applies with the choices $x_{ij_1j'_1} = 1/(d+1)$ for every vertex (i, j, i', j') of the graph. Under these conditions it follows hence that $\mathbf{P}(\bigcap_V A_{ij_1j'_1}^0) > 0$. Following in the footsteps of P. Erdős and J. Spencer⁷ we've obtained, with only a modest amount of effort, a crisp sufficient condition for the existence of Latin transversals.

THEOREM 1 *A square matrix A of order n has a Latin transversal if no element appears more than $(n-1)/(4e)$ times in A .*

The terminology for Latin transversals is of some antiquity and in this context a classical problem posed by Euler in 1782 is relevant. The problem asks for an arrangement of 36 officers of 6 ranks and from 6 regiments in a square formation of size 6 by 6. Each row and each column of this formation are to contain precisely one officer of each rank and precisely one officer from each regiment.

Euler's problem is related to the notion of what are called *Latin squares*. An $n \times n$ array with elements drawn from, say, the integers 1 through n forms a Latin square of order n if every row and every column contain n distinct integers. Latin squares may be thought of as multiplication tables for very general algebraic systems (fields). Readers familiar with the number game Sudoku will have realised that the valid number arrays in the game are just Latin squares with special additional structure.

Two Latin squares $A_1 = [a_{ij}^{(1)}]$ and $A_2 = [a_{ij}^{(2)}]$, each of order n , are said to be *orthogonal* if the n^2 pairs of elements $(a_{ij}^{(1)}, a_{ij}^{(2)})$ ($1 \leq i \leq n; 1 \leq j \leq n$) are all distinct. A pair of orthogonal Latin squares is also called a *Graeco-Latin square* or *Euler square*. Orthogonal Latin squares are closely linked to the notion of *finite projective planes* and have applications in coding theory.

Thus, $A_1 = [1]$ and $A_2 = [1]$ is the trivial pair of orthogonal Latin squares of order 1. It is not difficult to see that there are no orthogonal Latin squares of order 2; indeed, in this case, the Latin squares A_1 and A_2 have to be chosen from the two possibilities $\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$ and $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ and, for every selection, either the element pairs $(1, 1)$ and

⁷P. Erdős and J. Spencer, "Lopsided Lovász local lemma and Latin transversals", *Discrete Applied Mathematics*, vol. 14, pp. 151–154, 1990.

$(2, 2)$ cannot be recovered, or the element pairs $(1, 2)$ and $(2, 1)$ cannot be recovered. Orthogonal Latin squares of order 3 can be constructed, however, as shown below:

$$A_1 = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix}.$$

The combinatorial difficulties in construction get exponentially more severe as the order increases but a general result of the following stripe beckons.

THEOREM 2 *There exists a pair of orthogonal Latin squares of order n for every n satisfying $n \not\equiv 2 \pmod{4}$, that is to say, if n is not of the form $4k + 2$ for some integer k .*

The proof, while elementary, will take us a little afield; the interested reader will find a very readable account in H. J. Ryser's slim volume.⁸

What about the cases $n \equiv 2 \pmod{4}$? The case $n = 2$, as we have just seen, admits of no solution. The next case corresponds to $n = 6$ and, returning to our anecdote, we recognise Euler's problem as equivalent to the construction of a pair of orthogonal Latin squares of order 6. Euler conjectured that this was impossible and, indeed, G. Tarry verified this in 1900. Emboldened by this, we may perhaps be tempted to boldly conjecture, as Euler did, that there exists no pair of orthogonal Latin squares of order $n \equiv 2 \pmod{4}$. This was the subject of an extensive and involved sequence of investigations by R. C. Bose, S. S. Shrikhande, and E. T. Parker which culminated in 1960 with the demonstration of a remarkable theorem in an unexpected direction.⁹

THEOREM 3 *There exists a pair of orthogonal Latin squares of every order n satisfying $n \equiv 2 \pmod{4}$ and $n > 6$.*

Thus, the only cases with $n \equiv 2 \pmod{4}$ for which we do not have a pair of orthogonal Latin squares are when $n = 2$ or $n = 6$. This, as Ryser points out, illustrates the danger of leaping to general conclusions from limited empirical evidence. But we digress.

12 Problems

1. A closet contains ten distinct pairs of shoes. Four shoes are selected at random. What is the probability that there is at least one pair among the four selected?
2. *Coins.* A fair coin is tossed ten times. What is the probability that it falls heads at least three times in succession?
3. *Dice.* Five dice are thrown. Find the probability that at least three of them show the same face.

⁸H. J. Ryser, *Combinatorial Mathematics*. The Carus Mathematical Monographs, Mathematical Association of America, 1963.

⁹R. C. Bose, S. S. Shrikhande, and E. T. Parker, "Further results on the construction of mutually orthogonal Latin squares and the falsity of Euler's conjecture", *Canadian Journal of Mathematics*, vol. 12, pp. 189–203, 1960.

4. Suppose A_1, \dots, A_n are independent events, $P(A_j) = p_j$ for each j . Determine by inclusion-exclusion the probability that none of the A_j occur.
5. In the case of the Bose-Einstein statistics of Problem I.12 determine the probability that exactly m urns are empty by inclusion-exclusion.
6. *Derangements of rooks.* Eight mutually antagonistic rooks are placed randomly on a chessboard with all arrangements equally likely. Determine the probability that no rook can capture any other rook (that is to say, each rook occupies a distinct row and column).
7. *Generalisation.* A chessboard consists of n^2 squares arranged in n rows and n columns. Each square of the chessboard is identified by a pair of integers (i, j) with $1 \leq i, j \leq n$. A rook is a chess piece that when placed on square (i, j) can attack pieces on any square in the i th row or the j th column. Suppose n rooks are placed randomly on the chessboard. Determine the probability p_n that they are arranged in such a way that no two rooks can attack each other. Show that $p_n \sim 2\pi\sqrt{e}ne^{-2n}$ as $n \rightarrow \infty$. [Hint: You will need Stirling's formula: $n! \sim \sqrt{2\pi}n^{n+1/2}e^{-n}$.]
8. *Many dimensions.* In d dimensions the lattice points $\mathbf{i} = (i_1, \dots, i_d)$ where $1 \leq i_j \leq n$ may be identified with the "squares" (or, better perhaps, " d -dimensional unit cuboids") of a d -dimensional chessboard ranging over n cuboids in each dimension. A d -dimensional rook located at the lattice point (i_1, \dots, i_d) can range freely along points in directions parallel to the coordinate axes (varying i_j , for instance, while keeping i_k for $k \neq j$ fixed). Suppose r rooks are placed at random on the d -dimensional chessboard. Say that a given rook is "secure" if it does not lie along the axis-parallel lines of sight of any of the other $r-1$ rooks. Show that the number of secure rooks satisfies an asymptotic Poisson law for a critical rate of growth of $r = r_n$ with n .
9. *Permutations.* In a random permutation of the first n integers, by direct computation determine the probability that exactly r retain their original positions.
10. *Continuation.* In the previous problem set up a recurrence for the number of derangements d_n of the first n integers (that is, rearrangements with no integers in their original positions) by showing $d_{n+1} = nd_n + n d_{n-1}$ for $n \geq 2$ and thence solve for d_n .
11. *Ramsey numbers.* Show that $R(3, 3) = 6$. [Don't try your hand at $R(4, 4)$.]
12. *An upper bound for Ramsey numbers.* Show that $R(j, k) \leq \binom{j+k-2}{k-1}$. Hence show that $R(k, k) \leq 4^k / \sqrt{2\pi k}$.
13. *Permanents.* The permanent of a square matrix $A = [a_{ij}]$ whose order is n is defined by $\text{per}(A) = \sum a_{1j_1}a_{2j_2} \cdots a_{nj_n}$ where the sum extends over all permutations (j_1, j_2, \dots, j_n) of the integers $1, 2, \dots, n$. How many of the terms in the sum contain one or more diagonal elements of the matrix?
14. *Card collections.* Each box of a certain brand of cereal contains a randomly chosen baseball card from a collection of N distinct baseball cards. (The sampling is with replacement.) If $N = 3$, what is the smallest number of cereal boxes a collector will have to purchase to have a likelihood of greater than 50% of obtaining a complete set of baseball cards? Repeat for $N = 4$.
15. Let A_1, \dots, A_n be events with $P(A_j) = p$ for all j and $P(A_j \cap A_k) = q$ if $j \neq k$. You are told that it is certain that at least one of these events occur but that no more than two can occur. Show that $p \geq 1/n$ and $q \leq 2/n$.

16. *Bridge hand.* For $0 \leq m \leq 4$, determine the probability that a random bridge hand of 13 cards contains the ace–king pair of exactly m suits.

17. *Chromosome matching.* A cell contains n chromosomes between any two of which an interchange of parts may occur. If r interchanges occur [which can happen in $\binom{n}{2}^r$ ways] determine the probability that exactly m chromosomes are involved.¹⁰ (See Problem I.14 for another variation on the theme.)

18. *Matchings.* In a random shuffle of a deck of n distinguishable cards let $P_n(m)$ denote the probability that exactly m cards turn up in their original position. The weighted average $M_n = 0P_n(0) + 1P_n(1) + 2P_n(2) + \dots + nP_n(n)$ is the expected number of cards in their original position. Evaluate M_1 , M_2 , M_3 , and M_4 , corresponding to packs of one, two, three, and four cards, respectively. Proceed to find an asymptotic estimate for M_n when $n \rightarrow \infty$.

19. *Continuation.* Determine M_n for every n .

20. *Le problème des ménages.* At a dinner party, n couples are seated in a random arrangement at a circular table with men and women alternating. Show that the probability that no wife sits next to her husband is

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

This problem was formulated by E. Lucas. [Hint: It may be useful to show first that the number of ways of selecting k non-overlapping pairs of adjacent seats is given by $\binom{2n-k}{k} \frac{2n}{2n-k}$.]

21. *Partitions.* Show that the number of ways a set of n elements may be partitioned into k non-empty subsets is given by

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = \frac{1}{k!} \sum_{j=0}^k (-1)^j \binom{k}{j} (k-j)^n. \quad (12.1)$$

These are the Stirling numbers of the second kind.

22. *Connected graphs.* Let $g(n, m)$ be the number of connected graphs with n vertices and m edges. Let $f(n, m)$ be the number of these which have no vertices of degree one. Show that

$$f(n, m) = \sum_{k=0}^n (-1)^k \binom{n}{k} g(n-k, m-k) (n-k)^k \quad \text{if } n > 2.$$

(Observe that two vertices of degree one cannot be neighbours if the graph is connected.)

23. *Continuation, trees.* With notation as in (12.1), show that the number of trees on n vertices containing exactly r leaves is given by

$$T(n, r) = \frac{n!}{r!} \left\{ \begin{matrix} n-2 \\ n-r \end{matrix} \right\}.$$

¹⁰This was evaluated for $n = 6$ by D. G. Catcheside, D. E. Lea, and J. M. Thoday, "Types of chromosome structural change introduced by the irradiation of *tradescantia microspores*", *Journal of Genetics*, vol. 47, pp. 113–149, 1945.

24. Ten percent of the surface of a sphere is coloured blue; the rest of the surface is red. Show that it is always possible to inscribe a cube with vertices on the surface of the sphere such that all the vertices are red.

25. *Frequency-hopping collisions.* In a frequency-hopping spread spectrum¹¹ system for electrical communications the available frequency spectrum is divided into n non-overlapping bins (or subchannels) of equal bandwidth. The bandwidth of each subchannel is sufficiently large to accommodate user signalling requirements. Suppose there are m users in the system. Each user selects and uses a subchannel for a short period of time before moving (hopping) to a different subchannel. Suppose users select subchannels randomly and uniformly from the n available subchannels independently of each other and of previous selections. At any point in time, we say that a *collision* occurs if any two users occupy the same subchannel. Determine the probability $P(m, n)$ that there are no collisions.

26. *Continuation, asymptotics.* What is the largest rate of growth for $m = m_n$ with n for which $P(m_n, n)$ stays close to 1 as $n \rightarrow \infty$? [Recall that, for every fixed k , $-\log[(1 - 1/n)(1 - 2/n) \cdots (1 - k/n)] \sim k(k+1)/(2n)$ as $n \rightarrow \infty$.]

27. *Random triangles.* Write $G_{n,p}$ for the random graph on n vertices obtained by randomly and independently throwing each of the $\binom{n}{2}$ edges down with probability p . A collection of three vertices i, j , and k forms a triangle if the edges $\{i, j\}, \{i, k\}$, and $\{j, k\}$ are all in $G_{n,p}$. Determine the probability that there are exactly k triangles. What does the Poisson paradigm say here?

28. *4-cliques.* The random graph $G_{n,p}$ contains a clique on four vertices (or a K_4) if there exists a complete subgraph on four vertices. Suppose $p = p_n = c/n^{2/3}$. Show that the probability that $G_{n,p}$ contains no K_4 tends to $e^{-c^6/24}$ as $n \rightarrow \infty$.



29. *Off-diagonal Ramsey numbers.* For any $j \geq 2$ and $k \geq 2$, the Ramsey number $R(j, k)$ is the smallest integer n for which *any* two-colouring (say black and red) of the edges of K_n , the complete graph on n vertices, contains a black K_j or a red K_k . Use a probabilistic argument to determine an asymptotic lower bound for $R(3, k)$ for large values of k .

¹¹The discovery of frequency-hopping for secure communications is usually credited to an unexpected source. Facing the spectre of awful war on the eve of American entry into World War II and prompted by a desire to help her adopted country, the Austrian and Hollywood screen siren Hedy Kiesler Markey (better known by her marquee name Hedy Lamarr) improbably teamed up with the avant garde composer George Antheil and conceived of a novel system for use in warfare. Their “Secret Communications System” proposed frequency-hopping to make it difficult for the enemy to intercept or jam radio-guided torpedoes and was granted U.S. Patent 2,292,387 in 1942. Classified versions of the idea had previously also been used by the military on a limited scale to secure communications from enemy eavesdroppers. Out of barbarism sometimes emerges a rose of civilisation. Variations on the theme for commercial and civilian use spread in the latter part of the twentieth century with the proliferation of personal wireless communication devices; examples include the widely used code division multiple access systems for mobile telephony (the reader will find a sanitised analysis of a direct sequence, code division multiple access system which enables multiple users to communicate simultaneously over a common channel in Section VI.8) and the adaptive frequency-hopping spread spectrum systems used in the Bluetooth protocols for short-range communications between wireless devices.

Numbers Play a Game of Chance

Checking independence across events becomes more and more difficult as the number of events increases and the intuitive content of independence as a rule of products becomes harder to perceive. What then motivates the notion of independence as a rule of “multiplication of probabilities”, especially in infinite-dimensional cases where intuition may be totally abeyant? The key lies in E. Borel’s discovery in 1909 that numbers are “independent”. The fundamental underlying link between probability theory on the one hand and the theory of numbers on the other belongs to the earliest chapter of the history of probability and deserves a central and prominent rôle both from the point of view of clarifying the origins of the key idea of independence as well as for the rich cross-fertilisation of both fields that results. M. Kac has argued very eloquently for this point of view and we shall follow in his footsteps.¹

c 1-7

As we shall see, deep results in the theory of probability are already within our grasp via this correspondence with the theory of numbers: we shall see versions of two of the pillars of classical probability, the weak law of large numbers and the strong law of large numbers in this chapter; in the succeeding chapter we shall see the original version of the magisterial normal law, the *éminence grise* of probability. The benefits extend the other way to number theory as well; but illustrations in this direction will have to wait till Chapters XVI and XIX when we will have amassed a little more background. The material of this chapter and the next is not required elsewhere in this book (excepting only the dangerous bend sections XII.9 and XVI.10) and may be read independently.

1 A formula of Viète

We will embark on our historical pilgrimage by stepping back in time to a discovery of Viète and then proceeding via a wonderful structural detour to

¹M. Kac, *Statistical Independence in Probability, Analysis, and Number Theory*, the Carus Mathematical Monographs, Number 12, Mathematical Association of America, 1959.

Rademacher and thence to Borel. We begin with the elementary trigonometric identity $\sin 2\theta = 2 \sin \theta \cos \theta$ and, by repeated application, obtain the sequence of equations

$$\sin x = 2 \sin \frac{x}{2} \cos \frac{x}{2} = 2^2 \sin \frac{x}{4} \cos \frac{x}{4} \cos \frac{x}{2} = 2^3 \sin \frac{x}{8} \cos \frac{x}{8} \cos \frac{x}{4} \cos \frac{x}{2}.$$

Proceeding systematically in this fashion by induction we hence obtain the identity

$$\sin x = 2^n \sin \frac{x}{2^n} \prod_{k=1}^n \cos \frac{x}{2^k} \quad (1.1)$$

valid for each positive n . But by l'Hôpital's rule, if $x \neq 0$,

$$1 = \lim_{n \rightarrow \infty} \frac{\sin \frac{x}{2^n}}{\frac{x}{2^n}} = \frac{1}{x} \lim_{n \rightarrow \infty} 2^n \sin \frac{x}{2^n},$$

whence $\lim_n 2^n \sin \frac{x}{2^n} = x$. Dividing both sides of (1.1) by x and allowing n to tend to infinity, we hence obtain

$$\frac{\sin x}{x} = \prod_{k=1}^{\infty} \cos \left(\frac{x}{2^k} \right). \quad (1.1')$$

We may extend the identity to the case $x = 0$ as well if we identify both sides of the equation as unit.

This remarkable formula was discovered by François Viète in the second half of the sixteenth century. Though Viète's writings were elliptical, his discovery legitimised the consideration of formally infinite mathematical processes. In short order a number of other infinite processes—convergent products and series—were discovered, influencing the young Isaac Newton and leading in a direct line to the discovery of the binomial theorem and the calculus. Not bad for a little trigonometric identity. And if that were all there were to the story, it would still be a pretty thing to add to one's resume. But more was to follow: Viète's formula was to bear fruit in an unexpected direction three centuries later.

A NUMBER-THEORETIC ASIDE: CALCULATING π

The calculation of π to ever-increasing levels of precision is a spectator sport, the current record holder providing more than 10^{12} digits in the decimal expansion. While explorations of, and approximations to, π have been known since antiquity, approaches were geometric in nature until the discovery of infinite series and products. Formulae like those of Viète provide a mechanism that can compactly produce, in principle, any desired degree of approximation.

If we substitute $x = \pi/2$ in Viète's formula we obtain

$$\frac{2}{\pi} = \frac{\sqrt{2}}{2} \cdot \frac{\sqrt{2 + \sqrt{2}}}{2} \cdot \frac{\sqrt{2 + \sqrt{2 + \sqrt{2}}}}{2} \dots,$$

and, following in Viète's footsteps, we've discovered an unexpected infinite product representation for π . Viète discovered his remarkable formula in 1593. The formula, however, is not in itself particularly efficient as an estimator for π —to obtain each new term in the decimal expansion for π requires approximately two additional terms in the product. Here, for instance, is an evaluation of π from the product of the first 50 terms shown to 42 decimal places

$$\underline{3.141592653589793238462643383278483732551771 \dots}$$

compared to the actual value of π

$$\underline{3.141592653589793238462643383279502884197169 \dots}.$$

We've picked up 29 terms in the decimal expansion for π from 50 terms in Viète's product. It's cute but the convergence is glacial. Shortly thereafter John Wallis discovered several other infinite expansions that converge much faster and improvements came quickly as the pace of discovery quickened; of these, particularly potent representations were discovered by Srinivasa Ramanujan in the early part of the twentieth century.

2 Binary digits, Rademacher functions

Viète's formula was to have an unexpected consequence. To explore this more fully we will need to take a detour through number systems.

Every real number t in the unit interval $0 \leq t < 1$ has a *binary expansion* (or *dyadic expansion*)

$$t = \frac{z_1}{2} + \frac{z_2}{2^2} + \frac{z_3}{2^3} + \dots = .z_1 z_2 z_3 \dots \quad (\text{base 2}) \quad (2.1)$$

where each z_k is a binary digit, 0 or 1. The expansion is not necessarily unique. For instance, $5/8$ has the two representations $.101000\dots$ and $.100111\dots$, the first with a trailing infinite number of repeated 0s, the second with a trailing infinite number of repeated 1s:

$$\frac{5}{8} = \frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{0}{2^4} + \frac{0}{2^5} + \dots = \frac{1}{2} + \frac{0}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \frac{1}{2^5} + \dots.$$

As a *convention*, when there are two representations for t we agree to always use the terminating binary expansion, that is to say, the expansion with a trailing infinite number of zeros. Thus, we agree to use the first of the two expansions above, $5/8 = .1010$ rather than $5/8 = .10011$. With this convention in force, the binary expansion (2.1) is unique for every value of t .

Each binary digit z_k in the expansion (2.1) is completely determined by the real number t . In other words, $z_k = z_k(t)$ is a function of t and we may, more accurately, write (2.1) in the form

$$t = \sum_{k=1}^{\infty} \frac{z_k(t)}{2^k} \quad (0 \leq t < 1). \quad (2.1')$$

What do the functions $z_k(t)$ look like? A little introspection shows that the first three have the graphs shown in Figure 1. The gentle reader can easily

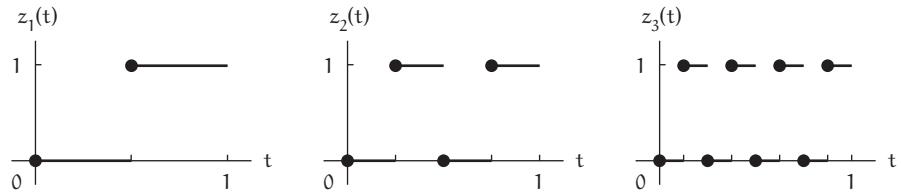


Figure 1: Binary digits.

discern the trend here. In general, $z_k(t)$ is a right continuous² step function which oscillates between the values of 0 and 1; the function is constant over intervals of length exactly 2^{-k} and exhibits jumps exactly at the points $j2^{-k}$ for $1 \leq j \leq 2^k - 1$.

For reasons of symmetry it is actually more convenient to consider the number $1 - 2t$ taking values in $(-1, +1]$. From (2.1') we obtain

$$1 - 2t = 1 - \sum_{k=1}^{\infty} \frac{2z_k(t)}{2^k} = \sum_{k=1}^{\infty} \frac{1 - 2z_k(t)}{2^k},$$

where we've exploited the fact that the geometric series $\sum_{k=1}^{\infty} 2^{-k}$ sums to 1. By setting

$$r_k(t) := 1 - 2z_k(t) \quad (k = 1, 2, \dots),$$

we then obtain a companion series for (2.1'),

$$1 - 2t = \sum_{k=1}^{\infty} \frac{r_k(t)}{2^k} \quad (0 \leq t < 1). \quad (2.1'')$$

The functions $r_k(t)$ were first studied by Hans A. Rademacher. They have several features of particular interest in the analytic theory of functions but for our purposes we consider them primarily in the rôle as surrogates of the binary digits in the dyadic expansions of real numbers.

The first three Rademacher functions are graphed in Figure 2. The trend is clear and it is not difficult now to provide a formal definition. For each $k = 1, 2, \dots$, partition the unit interval $[0, 1)$ into 2^k subintervals \mathbb{I}_{jk} each of length 2^{-k} and defined by

$$\mathbb{I}_{jk} = \left[\frac{j-1}{2^k}, \frac{j}{2^k} \right) \quad (j = 1, \dots, 2^k).$$

²The adoption of a non-terminating convention for the dyadic expansion would result in the same kind of picture except that the functions will now be continuous from the left.

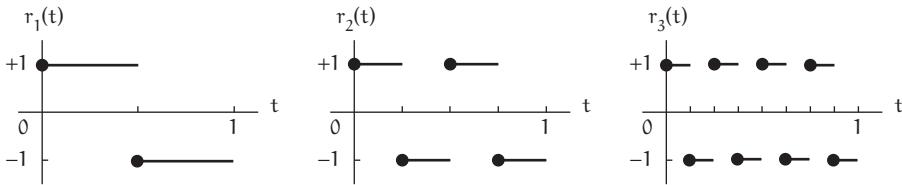


Figure 2: Rademacher functions.

The *Rademacher functions* $r_k(t)$ are then defined for each integer $k = 1, 2, \dots$, by

$$r_k(t) = \begin{cases} +1 & \text{if } t \text{ lies in } \mathbb{I}_{jk} \text{ with } j \text{ odd,} \\ -1 & \text{if } t \text{ lies in } \mathbb{I}_{jk} \text{ with } j \text{ even.} \end{cases}$$

A little thought shows that these functions can be specified recursively. Define the periodic function

$$r(t) = \begin{cases} +1 & \text{if } 0 \leq t \bmod 1 < 1/2, \\ -1 & \text{if } 1/2 \leq t \bmod 1 < 1, \end{cases}$$

where t varies over all real values. Here $t \bmod 1$ represents the fractional part of t .³ As base of a recurrence, set $r_1(t) = r(t)$ for $0 \leq t < 1$. The Rademacher functions then satisfy the recurrence

$$r_k(t) = r_{k-1}(2t \bmod 1) = r(2^{k-1}t) \quad (0 \leq t < 1)$$

for $k \geq 2$.

It is worth taking immediate stock of a key property of the Rademacher functions.

THEOREM 1 (ORTHOGONALITY) *For every integer $s \geq 1$, and every choice of distinct indices k_1, k_2, \dots, k_s , we have*

$$\int_0^1 r_{k_1}(t) r_{k_2}(t) \cdots r_{k_s}(t) dt = 0.$$

Or, in other words, the Rademacher functions are orthogonal.

The definition shows that $\int_0^1 r_k(t) dt = 0$ for each k , and an examination of the graphs of the functions makes clear why they are orthogonal. It is highly recommended that the reader provide her own proof.

³Every real number t can be expressed uniquely in the form $t = n + f$ where n is an integer and $0 \leq f < 1$. Then $t \bmod 1 = f$.

Before proceeding further, recall the cheeky formula for the complex exponential $e^{ix} = \cos x + i \sin x$ discovered by Euler. Here, as usual, $i = \sqrt{-1}$ is the imaginary root of unity. Immediate consequences are the trigonometric formulæ $\cos x = \frac{1}{2}(e^{ix} + e^{-ix})$ and $\sin x = \frac{1}{i2}(e^{ix} - e^{-ix})$.

To return to our discussion, the set of points t for which the Rademacher function $r_k(t)$ takes value $+1$ is comprised of 2^{k-1} subintervals, each of length 2^{-k} . It follows that the length (or *Lebesgue measure*) of the set of points where $r_k(t) = +1$ is exactly $2^{k-1}/2^k = 1/2$. Likewise, the length of the set of points where $r_k(t) = -1$ is also $1/2$. In consequence,

$$\int_0^1 e^{ixr_k(t)/2^k} dt = \frac{1}{2}e^{ix/2^k} + \frac{1}{2}e^{-ix/2^k} = \cos \frac{x}{2^k},$$

and it follows that the right-hand side of Viète's formula (1.1') is given by

$$\prod_{k=1}^{\infty} \cos \frac{x}{2^k} = \prod_{k=1}^{\infty} \int_0^1 \exp \left\{ \frac{ixr_k(t)}{2^k} \right\} dt.$$

On the other hand, an elementary exponential integration yields

$$\frac{\sin x}{x} = \int_0^1 e^{ix(1-2t)} dt = \int_0^1 \exp \left\{ \sum_{k=1}^{\infty} \frac{ixr_k(t)}{2^k} \right\} dt = \int_0^1 \prod_{k=1}^{\infty} \exp \left\{ \frac{ixr_k(t)}{2^k} \right\} dt$$

where we've deployed the expansion (2.1'') for $1-2t$ in terms of the Rademacher functions. We thus obtain the following striking restatement of Viète's formula in terms of Rademacher functions or, what is the same thing, binary digits.

THEOREM 2 (VIÈTE, REPRISE) *The Rademacher functions satisfy the identity*

$$\int_0^1 \prod_{k=1}^{\infty} \exp \left(ix \frac{r_k(t)}{2^k} \right) dt = \prod_{k=1}^{\infty} \int_0^1 \exp \left(ix \frac{r_k(t)}{2^k} \right) dt. \quad (2.2)$$

Here is a remarkable formula: an integral of an infinite product is an infinite product of integrals! [The reader will scarce have to be reminded that $\int fg$ is in general not to be expected to equal $(\int f)(\int g)$.] This is strangely reminiscent of the product of probabilities property characterising statistical independence though there is as yet no apparent link between number representations and probability. Is this just a fortuitous coincidence? Well, we certainly should not dismiss it without further investigation and a little more digging into the structure of numbers appears warranted.

3 The independence of the binary digits

For any interval $[a, b]$, let $\lambda[a, b] = b - a$, and if $[a_1, b_1], \dots, [a_n, b_n]$ is any finite collection of disjoint intervals, set $\lambda(\bigcup_{i=1}^n [a_i, b_i]) = \sum_{i=1}^n \lambda[a_i, b_i]$. Then

λ defines a positive-valued set function which associates to any finite union of disjoint intervals the sum of the lengths of those intervals. We call λ *Lebesgue measure on the unit interval*. (In Section XII.3 we shall see how to extend this idea of length to the family of Borel sets but for now the ordinary notion of length will serve.) In this notation our proof of Viète's formula in the previous section hinges upon the observation that

$$\lambda\{t : r_k(t) = +1\} = \lambda\{t : r_k(t) = -1\} = \frac{1}{2}$$

for each value of k . Consider now the set of points t in the unit interval on which $r_1(t) = +1$, $r_2(t) = -1$, and $r_3(t) = +1$, simultaneously. An examination of the graphs of the first three Rademacher functions shows that this set of points is identically the interval $[\frac{2}{8}, \frac{3}{8})$, from which it follows that the measure (or length) of this set of points is given by

$$\lambda\{t : r_1(t) = +1, r_2(t) = -1, r_3(t) = +1\} = \lambda\left[\frac{2}{8}, \frac{3}{8}\right) = \frac{1}{8}.$$

On the other hand, as we've seen

$$\lambda\{t : r_1(t) = +1\} = \lambda\{t : r_2(t) = -1\} = \lambda\{t : r_3(t) = +1\} = \frac{1}{2},$$

and we may restate our result in the form

$$\begin{aligned} \lambda\{t : r_1(t) = +1, r_2(t) = -1, r_3(t) = +1\} \\ = \lambda\{t : r_1(t) = +1\} \lambda\{t : r_2(t) = -1\} \lambda\{t : r_3(t) = +1\}. \end{aligned}$$

A general result now beckons.

THEOREM (INDEPENDENT DIGITS) *Let k_1, \dots, k_n be any finite collection of distinct indices. Then*

$$\begin{aligned} \lambda\{t : r_{k_1}(t) = \delta_{k_1}, \dots, r_{k_n}(t) = \delta_{k_n}\} \\ = \lambda\{t : r_{k_1}(t) = \delta_{k_1}\} \times \dots \times \lambda\{t : r_{k_n}(t) = \delta_{k_n}\} \quad (3.1) \end{aligned}$$

for every choice of constants $\delta_{k_1}, \dots, \delta_{k_n}$ taking values in $\{-1, +1\}$.

PROOF: It will suffice to show that

$$\lambda\{t : r_1(t) = \delta_1, \dots, r_n(t) = \delta_n\} = \lambda\{t : r_1(t) = \delta_1\} \times \dots \times \lambda\{t : r_n(t) = \delta_n\} \quad (3.2)$$

for every positive integer n and every choice of $\delta_1, \dots, \delta_n$. (Why?) As induction hypothesis, suppose that the set of points t on which $r_1(t) = \delta_1, \dots, r_n(t) = \delta_n$ is an interval of length 2^{-n} which, in turn, clearly implies the equality (3.2). This interval must then be of the form $[\frac{(j-1)}{2^n}, \frac{j}{2^n})$ for some integer j . [Why? The function $r_n(\cdot)$ changes sign in every dyadic interval of length 2^{-n} .] The

midpoint of the given interval partitions it into two contiguous subintervals of length $2^{-(n+1)}$ apiece with $r_{n+1}(t) = +1$ on the first of these subintervals and $r_{n+1}(t) = -1$ on the second. It follows that the set of points t on which $r_1(t) = \delta_1, \dots, r_n(t) = \delta_n$, and $r_{n+1}(t) = \delta_{n+1}$ is one of these two subintervals of length $2^{-(n+1)}$. The induction and the proof are complete. ▶

The Rademacher functions satisfy a product law akin to independence! While I have chosen to phrase the result in terms of the Rademacher functions $\{r_k(t), k \geq 1\}$, the reader should realise that we may equally well have phrased the product property (3.1) in terms of the binary digits $\{z_k(t), k \geq 1\}$ and we say hence that *the binary digits are independent*.

On the surface (3.2) might seem a trivial restatement of the obvious identity

$$\frac{1}{2^n} = \underbrace{\frac{1}{2} \times \cdots \times \frac{1}{2}}_{n \text{ times}},$$

but as E. Borel discovered in 1909 it is much much more. This is the key to (2.2).

ANOTHER PROOF OF VIÈTE'S FORMULA: Write $c_k = x2^{-k}$ and consider the sum $\sum_{k=1}^n c_k r_k(t)$. As t varies over the unit interval, the Rademacher functions $r_1(t), \dots, r_n(t)$ each take values only in $\{-1, +1\}$ and the sum $\sum_{k=1}^n c_k r_k(t)$ can hence take on only one of 2^n possible distinct values, each of the form $\pm c_1 \pm c_2 \pm \cdots \pm c_n$. Thus, as t varies over the unit interval, the vector $r(t) = (r_1(t), \dots, r_n(t))$ runs through the set of 2^n values $\delta = (\delta_1, \dots, \delta_n) \in \{-1, +1\}^n$ and the sum $\sum_{k=1}^n c_k r_k(t)$ consequently runs through the values $\sum_{k=1}^n c_k \delta_k$. Hence

$$\begin{aligned} \int_0^1 \prod_{k=1}^n \exp\{ic_k r_k(t)\} dt &= \int_0^1 \exp\left\{i \sum_{k=1}^n c_k r_k(t)\right\} dt \\ &= \sum_{\delta \in \{-1, +1\}^n} \exp\left\{i \sum_{k=1}^n c_k \delta_k\right\} \lambda\{t : r(t) = \delta\}. \end{aligned} \quad (3.3)$$

The independence of the binary digits yields

$$\lambda\{t : r(t) = \delta\} = \lambda\{t : r_1(t) = \delta_1\} \times \cdots \times \lambda\{t : r_n(t) = \delta_n\},$$

with another product over n terms obtaining from the product property of the exponential, $e^{i(c_1 \delta_1 + \cdots + c_n \delta_n)} = e^{ic_1 \delta_1} \times \cdots \times e^{ic_n \delta_n}$. Grouping terms in the sum on the right of (3.3) yields

$$\begin{aligned} &\sum_{\delta_1, \dots, \delta_n} [e^{ic_1 \delta_1} \lambda\{t : r_1(t) = \delta_1\}] \times \cdots \times [e^{ic_n \delta_n} \lambda\{t : r_n(t) = \delta_n\}] \\ &= \left\{ \sum_{\delta_1} e^{ic_1 \delta_1} \lambda\{t : r_1(t) = \delta_1\} \right\} \times \cdots \times \left\{ \sum_{\delta_n} e^{ic_n \delta_n} \lambda\{t : r_n(t) = \delta_n\} \right\}. \end{aligned}$$

Replacing c_k by $x/2^k$, we hence obtain

$$\int_0^1 \prod_{k=1}^n \exp\left(ix \frac{r_k(t)}{2^k}\right) dt = \prod_{k=1}^n \int_0^1 \exp\left(ix \frac{r_k(t)}{2^k}\right) dt. \quad (3.4)$$

It only remains to take the limit as $n \rightarrow \infty$ of both sides. On the right we then immediately obtain the infinite product of integrals in Viète's formula, while on the left we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \int_0^1 \prod_{k=1}^n \exp\left(ix \frac{r_k(t)}{2^k}\right) dt &= \int_0^1 \lim_{n \rightarrow \infty} \prod_{k=1}^n \exp\left(ix \frac{r_k(t)}{2^k}\right) dt \\ &= \int_0^1 \prod_{k=1}^{\infty} \exp\left(ix \frac{r_k(t)}{2^k}\right) dt, \end{aligned}$$

the interchange of the operations of limit and integral permissible as a consequence of the uniform convergence of the series $\sum_{k=1}^{\infty} r_k(t)2^{-k}$ to $1 - 2t$. ►

The reader familiar with the notion of uniform convergence will need no further convincing but the reader whose memory is a little rusty may find an excursus on convergence useful. Suppose $\{f_n, n \geq 1\}$ is a sequence of functions on the unit interval. Then f_n converges uniformly to a limit function f if, for every $\epsilon > 0$, there exists $n_0 = n_0(\epsilon)$ so that $|f_n(t) - f(t)| < \epsilon$ for every t whenever $n \geq n_0$. (What makes the convergence uniform is that the choice of $n \geq n_0$ sufficiently large is determined solely by the desired degree of approximation ϵ and is independent of the argument t .) If we identify $f_n(t) := \sum_{k=1}^n x r_k(t) 2^{-k}$ and $f(t) := \sum_{k=1}^{\infty} x r_k(t) 2^{-k}$ then $|f_n(t) - f(t)| \leq |x| \sum_{k=n+1}^{\infty} 2^{-k} = |x| 2^{-n}$ and *a fortiori* f_n converges uniformly to f on the unit interval. To see that this suffices to permit us to exchange limit and integral, consider

$$\left| \int_0^1 e^{if(t)} dt - \int_0^1 e^{if_n(t)} dt \right| = \left| \int_0^1 e^{if(t)} [1 - e^{i\{f_n(t) - f(t)\}}] dt \right| \leq \int_0^1 |1 - e^{i\{f_n(t) - f(t)\}}| dt,$$

the upper bound following because $e^{if(t)}$ has unit modulus. The integrand on the right may be expressed in the explicit trigonometric form

$$\begin{aligned} |1 - e^{i\{f_n(t) - f(t)\}}| &= \sqrt{\{1 - \cos(f_n(t) - f(t))\}^2 + \sin(f_n(t) - f(t))^2} \\ &= \sqrt{2 - 2 \cos(f_n(t) - f(t))} = 2 |\sin(\frac{f_n(t) - f(t)}{2})| \end{aligned}$$

courtesy the trigonometric identity $2 \sin(\theta)^2 = 1 - \cos(2\theta)$ deployed at the last. Now select any small, strictly positive ϵ . Then for all sufficiently large n we have $|f_n(t) - f(t)| < \epsilon$ uniformly for all t . Accordingly,

$$|1 - e^{i\{f_n(t) - f(t)\}}| < 2 \sin\left(\frac{\epsilon}{2}\right) \leq \epsilon$$

in view of the elementary result $\sin \theta \leq \theta$ for all $\theta \geq 0$. This last observation is an immediate consequence of the mean value theorem: there exists $0 \leq \theta_1 \leq \theta$ so that

$\sin(\theta) = \theta \sin'(\theta_1) = \theta \cos(\theta_1)$. As $\cos(\theta_1)$ is bounded above by 1, it follows that $\sin(\theta) \leq \theta$ for all positive θ . We've hence shown that

$$\left| \int_0^1 e^{if(t)} dt - \int_0^1 e^{if_n(t)} dt \right| < \epsilon$$

for all sufficiently large n . As $\epsilon > 0$ may be chosen arbitrarily small, it follows that the exchange of limit and integral is indeed permissible as asserted.

It is a matter of taste which proof of Viète's formula one prefers. Our elegant first proof is essentially trigonometric: Viète's formula drops into our lap almost serendipitously through manipulations of Euler's trigonometric formulæ. Our second proof is essentially analytical in nature and, in contrast, lays out the machinery for inspection. This proof makes apparent that it is the property (3.1) of the binary digits that is at the heart of Viète's formula. The product property of independent events bears a striking resemblance to the independence property of the binary digits and a path to *modelling* probability experiments via number systems now suggests itself.

4 The link to coin tossing

We have already explored the connection between coin-tossing and the real numbers in Examples I.7.6, 7. Here is the correspondence, made explicit.

COIN TOSSES

An unlimited sequence of tosses of a fair coin consists of a sequence of symbols $\omega_1, \omega_2, \dots, \omega_n, \dots$ where each symbol takes one of two values, \mathfrak{H} or \mathfrak{T} . The events A of interest are infinite collections of such sequences and may be specified in terms of *cylinder sets* where each cylinder set is a family of all sequences $\omega = \{\omega_k, k \geq 1\}$ satisfying conditions of the form

$$\omega_{k_1} = \alpha_{k_1}, \omega_{k_2} = \alpha_{k_2}, \dots, \omega_{k_n} = \alpha_{k_n}$$

for some finite selection of trials k_1, k_2, \dots, k_n , and a fixed selection of toss outcomes $\alpha_{k_1}, \alpha_{k_2}, \dots, \alpha_{k_n}$ taking values in $\{\mathfrak{H}, \mathfrak{T}\}$. In other words, a cylinder set is a collection of infinite sequences of coin tosses with the values of n of the tosses specified for some n . The probabilities of any such event A are most conveniently calculated by exploitation of the condition of "independent tosses" of the coin whereby, for each cylinder set of the above form,

$$\begin{aligned} P\{\omega_{k_1} = \alpha_{k_1}, \omega_{k_2} = \alpha_{k_2}, \dots, \omega_{k_n} = \alpha_{k_n}\} \\ = P\{\omega_{k_1} = \alpha_{k_1}\} \times P\{\omega_{k_2} = \alpha_{k_2}\} \times \dots \times P\{\omega_{k_n} = \alpha_{k_n}\} = \left(\frac{1}{2}\right)^n, \end{aligned}$$

as the occurrence of the event $\{\omega_{k_j} = \alpha_{k_j}\}$ depends only on the outcome of the k_j th toss and is independent of the other tosses.

BINARY DIGITS

Each sample point in the coin-tossing experiment is in one-to-one correspondence with a real number $t = .z_1 z_2 \dots$ in the unit interval determined by the binary digits $\{z_k(t), k \geq 1\}$ or, equivalently, the Rademacher functions $\{r_k(t), k \geq 1\}$. The result of the k th toss is now naturally enough identified with the value taken by the k th binary digit $z_k(t)$, the outcomes \mathfrak{H} and \mathfrak{T} for the toss identified with the binary digits 1 and 0, respectively.⁴ *The key to the correspondence is that each cylinder set $\{\omega_1 = \alpha_1, \omega_2 = \alpha_2, \dots, \omega_n = \alpha_n\}$ is in one-to-one correspondence with a unique interval of length $(\frac{1}{2})^n$.* Unions of cylinder sets are identified with unions of intervals and the requirement of consistency now mandates that each event of interest [that is to say, a subset A of sample points $\omega = (\omega_k, k \geq 1)$] in the coin-tossing problem is identified with a subset \mathbb{A} of points t in the unit interval. The probability of each such event A comprised of collections of toss sequences is then identified with the measure of the corresponding subset \mathbb{A} of points in the unit interval. These correspondences set up our “dictionary of translation”: *temporarily setting native coin-tossing intuition aside, we systematically assign to each event A in the sample space of repeated coin tosses a probability equal to the Lebesgue measure of the corresponding set \mathbb{A} of points in the unit interval.* The basic items in this correspondence between coin tosses and the binary digits are summarised in Table 1.

Terminology	Coin tosses	Binary digits
k th trial	$\omega_k \in \{\mathfrak{T}, \mathfrak{H}\}$	$z_k(t) \in \{0, 1\}$ or $r_k(t) \in \{+1, -1\}$
Sample point	$\omega = (\omega_k, k \geq 1)$	$t = \sum_k z_k(t) 2^{-k} = \frac{1}{2} - \frac{1}{2} \sum_k r_k(t) 2^{-k}$
Basic event	Cylinder set: I	Interval: \mathbb{I}
Event	Union of cylinder sets: A	Union of intervals: \mathbb{A}
Measure	Event probability: $P(A)$	Lebesgue measure: $\lambda(\mathbb{A})$

Table 1: A dictionary of translation modelling the repeated toss of a coin.

It should be kept in mind that what we have done here is provide a *model* for an idealised coin-tossing experiment in terms of the binary digits. The model provides an unambiguous procedure for computing event probabilities in the coin-tossing experiment: merely set the probability equal to the length, that is to say, the Lebesgue measure, of the corresponding set of points in the unit interval. While the *validity* of the model cannot, of course, be established mathematically, our reason for believing in it is because *the independence of the binary digits yields cylinder set probabilities that are consistent with empirical evidence from finite coin-tossing experiments.* The reader may recall that the extension of Lebesgue measure from interval lengths to general sets (the so-called Borel sets)

⁴Of course, the convention is arbitrary; we could equally well have identified \mathfrak{H} with 0 and \mathfrak{T} with 1.

is unique. (If the reader does not know this result she will find it subsumed in the general proof of uniqueness provided in Chapter XI.) *The uniqueness of Lebesgue measure guarantees hence that the probability measure induced by our model of coin tossing is unique.*

While Viète's formula relied upon a very particular additional property of the Rademacher functions, viz., the uniformly convergent series $1 - 2t = \sum_{k=1}^{\infty} \frac{r_k(t)}{2^k}$ (the dyadic expansion of a real number!), the fundamental property that drove the derivation of the result was the key discovery by Borel that the binary digits are independent. The statistical property of independent coin tosses is now seen to be simply the independence of the binary digits in another guise. *It is (3.1) then which is used as a jump-off point in the formal definition of "statistical independence" in the abstract theory of probability.* The critical link to our intuitive, experience-driven notion of "independence" (without which the theory would be sterile and bereft of the lifeblood provided by vibrant application) is provided by the correspondence with coin tossing.

Quo vadis? A natural generalisation is to consider how one might alter the dictionary of translation to take into account tossing an unfair coin whose success probability is p . The proper correspondence is then with a "bent" binary digit $z_k(t; p)$ which takes value 0 over a fraction p of the unit interval and value 1 over the remaining fraction $1 - p$ of the interval. The reader will find a sketch of how one might go about doing this in the *Problems* at the end of this chapter.

5 The binomial makes an appearance

In our dictionary of translation we have (temporarily) set coin-tossing intuition aside when going from a finite number of tosses to an infinite number of tosses and identified probabilities with Lebesgue measure. It is always wise when moving to an abstract framework to test the results one obtains in simple cases against intuition. Accordingly, to check out the bona fides of our dictionary of translation consider the following familiar question: what is the probability $b_n(k)$ that n tosses of a fair coin result in exactly k heads? In terms of our dictionary, this appears in a new guise: what is the measure of the set of points t in the unit interval for which

$$z_1(t) + \cdots + z_n(t) = k \quad \text{or, equivalently,} \quad r_1(t) + \cdots + r_n(t) = n - 2k? \quad (5.1)$$

Of course, there is a direct combinatorial route to the answer, but the object of this exercise is to test the bona fides of the dictionary of translation that we have set up. Via our correspondence between coin tosses and binary digits, we have

$$b_n(k) = \lambda\{t : r_1(t) + r_2(t) + \cdots + r_n(t) = n - 2k\} = \int_{\mathbb{A}} dt = \int_0^1 1_{\mathbb{A}}(t) dt$$

where $\mathbb{A} = \mathbb{A}(k, n)$ is the set of points t in the unit interval for which condition (5.1) is satisfied and $1_{\mathbb{A}}(t)$ is the *indicator function* for \mathbb{A} defined by

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

Now, wouldn't it be good if we could find a nice analytical expression for \mathbb{A} or its indicator $1_{\mathbb{A}}(t)$? We start with the easy observation that for integer κ ,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ix\kappa} dx = \begin{cases} 1 & \text{if } \kappa = 0, \\ 0 & \text{if } \kappa \neq 0. \end{cases}$$

Writing $R_n(t) = r_1(t) + \dots + r_n(t)$ for brevity, the indicator function for \mathbb{A} can then be written in the form

$$1_{\mathbb{A}}(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ix\{R_n(t)-(n-2k)\}} dx.$$

This is clever: we've managed to link the mysterious set \mathbb{A} to an expression involving the Rademacher functions. The independence of the binary digits mops things up:

$$\begin{aligned} \int_0^1 \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ix\{R_n(t)-(n-2k)\}} dx \right\} dt &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ix(n-2k)} \left\{ \int_0^1 \prod_{j=1}^n e^{ixr_j(t)} dt \right\} dx \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ix(n-2k)} \prod_{j=1}^n \left\{ \int_0^1 e^{ixr_j(t)} dt \right\} dx. \end{aligned}$$

The interchange of integrals in the first step is easily seen to be permissible as the integrand is absolutely integrable while the second step is just the independence property of the Rademacher functions. As $\int_0^1 e^{ixr_j(t)} dt = \cos(x)$ for each j , we obtain

$$b_n(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ix(n-2k)} \cos(x)^n dx. \quad (5.2)$$

So, we are left with an elementary integral to evaluate. How do we go about doing it? At least three approaches suggest themselves.

Induction: the direct approach. Two integrations by parts yield the basic recurrence

$$b_n(k) = \frac{n(n-1)}{4k(n-k)} b_{n-2}(k-1) \quad (1 \leq k \leq n-1; n \geq 2)$$

with boundary conditions

$$b_n(k) = 2^{-n} \quad (\text{if } k = 0 \text{ or } k = n \text{ for all } n \geq 0).$$

Euler's formula: the quickest approach. The familiar trigonometric identity of Euler coupled with the binomial theorem yields

$$\cos(x)^n = \left(\frac{e^{ix} + e^{-ix}}{2} \right)^n = 2^{-n} \sum_{j=0}^n \binom{n}{j} e^{i(n-j)x} e^{-ijx}.$$

Which immediately leads to

$$b_n(k) = 2^{-n} \sum_{j=0}^n \binom{n}{j} \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i2x(k-j)} dx.$$

Fourier series: the slickest approach. The Fourier coefficients f_k (integer k) of the periodic function $\cos x$ are given by

$$f_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikx} \cos(x) dx = \begin{cases} \frac{1}{2} & \text{if } k = \pm 1, \\ 0 & \text{otherwise.} \end{cases}$$

The Fourier coefficients of the periodic function

$$\cos(x)^n = \underbrace{\cos x \times \cos x \times \cdots \times \cos x}_{n \text{ times}}$$

are hence

$$\{f_k^{(n)}\} = \underbrace{\{f_k\} * \{f_k\} * \cdots * \{f_k\}}_{n\text{-fold convolution}}.$$

(The convolution property of Fourier series!) The reader will observe that the Fourier coefficients build up via Pascal's triangle and she can finish up with the observation

$$b_n(k) = f_{n-2k}^{(n)}.$$

Any of the above approaches yields the desired expression:

$$b_n(k) = \binom{n}{k} 2^{-n} \quad (0 \leq k \leq n; n \geq 0). \quad (5.2')$$

For every n , the terms $b_n(k)$ ($0 \leq k \leq n$) determine a discrete probability distribution. This is the famous *binomial distribution*.

A reader may be forgiven for not being very impressed; there is after all a direct combinatorial path to the answer. By enumerating the positions of the heads, the number of sequences of n tosses that result in exactly k heads is $\binom{n}{k}$. And (5.2') follows as any given realisation of n tosses has probability 2^{-n} . While it is true that the combinatorial approach is much simpler here, the analytical approach through the Rademacher functions pays dividends by uncovering a bridge between areas which appear on the surface to be quite unconnected. A pathway has now been created to press into service the tools and intuition peculiar to the theory of numbers into probabilistic investigations.

6 An inequality of Chebyshev

A question of importance to a gambler: what is the probability that 100 tosses of a fair coin result in more than 60 successes? Or in fewer than 40 successes? This is an example of a *tail probability* as the events of interest lie in the tail of the distribution. More generally, let S_n be the number of heads seen in n tosses of a fair coin. It is important in many applications to be able to estimate probabilities of the form $P\{|S_n - n/2| \geq \epsilon n\}$, the probability that the number of heads differs from $n/2$ in absolute value by ϵn or more; in these settings, $\epsilon > 0$ is small, n is large, and these tail events usually connote rare, but important events.

In consequence of (5.2'), we can directly write

$$P\{|S_n - n/2| \geq \epsilon n\} = \sum_{k:|k-n/2| \geq \epsilon n} b_n(k) = 2^{-n} \sum_{k:|k-n/2| \geq \epsilon n} \binom{n}{k}. \quad (6.1)$$

While the expression on the right, unfortunately, does not in general have a closed form, for large values of n it is possible to approximate the sum by first estimating the asymptotic behaviour of the summands. To see how an approach like this may unfold, the reader is invited to find some fun and profit in attempting to estimate the binomial coefficients directly using Stirling's formula for the factorial, $n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n}$. (If the reader does not know it she will find a proof in Section XIV.7.)

While a calculation based on a direct approach to estimating the binomial coefficients has the great virtue of being elementary, it has the disadvantage that it is particularly crafted to take advantage of the peccadilloes of the binomial and it is not immediately obvious how to generalise it. Let us see what an approach based upon the correspondence of coin tossing with the binary digits brings up.

The event that there are k heads in n tosses corresponds to the set of points t in the unit interval satisfying

$$z_1(t) + \cdots + z_n(t) = k \quad \text{or, equivalently,} \quad r_1(t) + \cdots + r_n(t) = n - 2k.$$

With $R_n(t) = r_1(t) + \cdots + r_n(t)$ as before, the event $\{|S_n - n/2| \geq \epsilon n\}$ corresponds then via our dictionary of translation to the set of points

$$\mathbb{A} = \mathbb{A}_n(\epsilon) = \{ t : |R_n(t)| \geq 2\epsilon n \}.$$

It follows that the desired event probability may be identified with the measure of \mathbb{A} :

$$P\{|S_n - n/2| \geq \epsilon n\} = \lambda\{ t : |R_n(t)| \geq 2\epsilon n \} = \int_{\mathbb{A}} dt = \int_0^1 1_{\mathbb{A}}(t) dt,$$

where, as before, $1_{\mathbb{A}}(t)$ is the indicator function for the set \mathbb{A} .

It is tempting to try the integration manoeuvre of the previous section but the reader will find that she will only have been successful in recovering (6.1) and we appear to be at an impasse. Consider, however, the following artifice introduced by Pafnuty Chebyshev. As t ranges over the set \mathbb{A} ,

$$\left(\frac{r_1(t) + \cdots + r_n(t)}{2\epsilon n} \right)^2 \geq 1 = 1_{\mathbb{A}}(t) \quad (t \in \mathbb{A}),$$

while outside \mathbb{A} ,

$$\left(\frac{r_1(t) + \cdots + r_n(t)}{2\epsilon n} \right)^2 \geq 0 = 1_{\mathbb{A}}(t) \quad (t \notin \mathbb{A}).$$

Thus the indicator for \mathbb{A} is dominated by

$$1_{\mathbb{A}}(t) \leq \left(\frac{r_1(t) + \cdots + r_n(t)}{2\epsilon n} \right)^2$$

for all t in the unit interval. It follows that

$$\mathbf{P}\{|S_n - n/2| \geq \epsilon n\} \leq \int_0^1 \left(\frac{r_1(t) + \cdots + r_n(t)}{2\epsilon n} \right)^2 dt.$$

Expanding the square inside the integral via the multinomial theorem, we are left with integrals of the form $\int_0^1 r_k(t)^2 dt$ and $\int_0^1 r_j(t)r_k(t) dt$. As $r_k(t)^2 = 1$ for every k , the former integrals evaluate to 1. The orthogonality property of the Rademacher functions (Theorem 2.1) on the other hand shows that all integrals involving cross-product terms disappear. Thus,

$$\mathbf{P}\{|S_n - n/2| \geq \epsilon n\} \leq \frac{n}{4\epsilon^2 n^2} = \frac{1}{4\epsilon^2 n}. \quad (6.2)$$

As we shall see, the appearance of $\epsilon^2 n$ in the denominator is characteristic of the argument in much more general settings. Replacing ϵ by $\epsilon/2$ cleans up the expression a little without changing the essence of the argument. And *a fortiori* we have proved

THE WEAK LAW OF LARGE NUMBERS FOR COIN TOSSES *Suppose ϵ is any fixed, strictly positive quantity. Then*

$$\mathbf{P}\left\{\left|\frac{1}{n}S_n - \frac{1}{2}\right| \geq \epsilon\right\} \rightarrow 0 \quad (6.3)$$

as $n \rightarrow \infty$.

In a long sequence of coin tosses, the fraction of heads is very likely to be close to one-half. This statement is expressed compactly by saying that *the sequence $\frac{1}{n}S_n$*

converges in probability to 1/2, or, simply, $\frac{1}{n}S_n \rightarrow^p \frac{1}{2}$. The reader should resist the temptation to interpret this in the light of the ordinary notion of convergence of sequences. Convergence in probability refers to a much weaker phenomenon which, in this context, is just a short form for saying (6.3).

The binomial tail probability hence tends to zero. The rate at which the upper bound approaches zero is stodgy but nonetheless suffices unto the purpose! More importantly, this device of Chebyshev proves to be wildly useful.

Writing $P_n(\epsilon) = P\{|S_n - n/2| \geq \epsilon n\}$ to focus on the key elements, the question we posed at the start of this section asks us to estimate $P_{100}(0.1)$. In the bound (6.2) this corresponds to $n = 100$ and $\epsilon = 0.1$ and we obtain $P_{100}(0.1) \leq 0.25$. This is not a very good bound; the exact answer is $P_{100}(0.1) = 0.056\dots$. With the same ϵ , matters improve somewhat with increasing n as Chebyshev's bound decreases with n : the bound gives $P_{1000}(0.1) \leq 0.025$ (the actual value is about 2.73×10^{-10}) and the fact that the estimate yields a numerically small value may be good enough for some applications. The true worth of Chebyshev's inequality is not in that it provides accurate estimates (it doesn't) but that it provides a wonderfully simple tool of great versatility. Only experience will convince the casual reader that simple conditions like Chebyshev's are frequently much more useful than condition-beset theorems of apparent power. The applications in Chapters XVI and XVII may serve to convince the sceptical student.

Cogito ergo sum. The reader may wonder what inspired the bound of the indicator for A by $R_n(t)^2/4\epsilon^2 n^2$. She will get a little forrander if she considers how the argument would unfold if she had used

$$f_n(t) = \left(\frac{r_1(t) + \dots + r_n(t)}{2\epsilon n} \right)^4$$

instead in the Chebyshev argument. Or, more generally,

$$f_n(t) = \left(\frac{r_1(t) + \dots + r_n(t)}{2\epsilon n} \right)^M$$

for any choice of even positive integer M .

7 Borel discovers numbers are normal

The weak law says that, for a large enough value of n , the set of points t for which $\left| \frac{1}{n} (r_1(t) + \dots + r_n(t)) \right| \geq 2\epsilon$, or, equivalently, $\left| \frac{1}{n} (z_1(t) + \dots + z_n(t)) - \frac{1}{2} \right| \geq \epsilon$, has small (Lebesgue) measure. In other words, 0s and 1s are likely to be approximately evenly balanced up to n digits in a binary expansion. This does not imply, however, that a number t with evenly balanced digits up to a given point n will continue to exhibit balance in its digits beyond that point.

For example, the number $.010101\dot{0}$ has 0s and 1s balanced through $n = 6$ but 0s dominate from that point onwards. What can be said about the family of numbers for which 0s and 1s occur in fair (that is to say, equal) proportion in the *entire* dyadic expansion? This is the family of *normal numbers* in the terminology introduced by Borel. Let us formalise this notion.

Consider any number t in the unit interval $[0, 1)$ and its dyadic expansion $.z_1(t)z_2(t)\dots$ (base 2). The fraction of 1s that are present in the first n bits of the dyadic expansion is then given by

$$\frac{\text{\# of 1s in the first } n \text{ bits}}{n} = \frac{z_1(t) + \dots + z_n(t)}{n}.$$

Define the *density of 1s in the dyadic expansion of t* by

$$v(t) = \lim_{n \rightarrow \infty} \frac{z_1(t) + \dots + z_n(t)}{n}$$

if the limit exists.⁵ If $v(t)$ exists, we may then identify it with the long-run proportion of 1s in the dyadic expansion of t . Likewise, the *density of 0s in the dyadic expansion of t* may be identified with the long-run proportion of 0s in the expansion, again with the proviso that the limit exists. The reader will observe that if either limit exists then both do, their sum being identically one.

DEFINITION 1 We say a number $t \in [0, 1)$ is *normal in base 2* if the density $v(t)$ exists and is equal to $1/2$.

Thus, a number is normal if the long-run proportion of 1s is equal to the long-run proportion of 0s in its dyadic expansion. We are interested in the measure of the set $\{t \in [0, 1) : t \text{ is normal}\}$.

We begin with a little measure-theoretic detour to pick up one definition and one technical result.

DEFINITION 2 Any subset \mathbb{A} of the unit interval whose length is zero, i.e., $\lambda(\mathbb{A}) = 0$, is said to have (*Lebesgue*) *measure zero*. A property which holds for all points t in the unit interval with the exception of a set of points of measure zero is said to hold *almost everywhere* (usually abbreviated *a.e.*).

It is clear from elementary calculus that a single point $\{x_i\}$ or a finite collection of points $\{x_1, \dots, x_n\}$ or even a countable collection of points $\{x_i, i \geq 1\}$ has length zero. But much more complicated sets of measure zero can be constructed. The reader will see some examples in Section XI.3 where she will also find a more complete and satisfying discussion of the concept of length and how it may

⁵It is a sad fact, but true, that the limit may take any value between 0 and 1 and, indeed, may not even exist. It is instructive to construct an example.

be naturally extended to sets much more complicated than intervals and finite groupings of them.

The reader no doubt knows that interchanging the order of an infinite sum and an integral may be fraught with danger. Much of the great success of a unified theory of measure was in making such operations transparent. The following result provides a representative and useful example.

LEVI'S THEOREM *Suppose $\{f_n(t), n \geq 1\}$ is a sequence of positive functions defined on the unit interval $0 \leq t < 1$. If $\sum_{n=1}^{\infty} \int_0^1 f_n(t) dt$ is convergent then the series $\sum_{n=1}^{\infty} f_n(t)$ converges a.e.*

This is a well-known result due to Beppo Levi in the theory of integration. If the reader has not seen it before she will find it (and more besides) in the proof of Theorem XIII.5.3 and, again, in Section XIII.8.

We are now ready to begin an assault on normal numbers. Consider the inspired choice of positive functions

$$f_n(t) = \left(\frac{r_1(t) + \cdots + r_n(t)}{n} \right)^4 \quad (7.1)$$

for use in Levi's theorem. As $r_k(t)^2 = r_k(t)^4 = 1$ and $r_k(t)^3 = r_k(t)$, the multinomial theorem in conjunction with the orthogonality of the Rademacher functions yields

$$\int_0^1 f_n(t) dt = \frac{\binom{n}{1} + \binom{n}{2} \binom{4}{2}}{n^4} = \frac{3}{n^2} \left(1 - \frac{2}{3n} \right) < \frac{3}{n^2}$$

as the expansion of $f_n(t)$ yields precisely $\binom{n}{1}$ terms of the form $r_k(t)^4$ and $\binom{n}{2} \binom{4}{2}$ terms of the form $r_j(t)^2 r_k(t)^2$, all other cross-product terms integrating to zero. It follows that

$$\sum_{n=1}^{\infty} \int_0^1 f_n(t) dt < \sum_{n=1}^{\infty} \frac{3}{n^2} = 3 + \sum_{n=2}^{\infty} \frac{3}{n^2} < 3 + \int_1^{\infty} \frac{3}{x^2} dx = 6.$$

(The bound can be improved; for instance, $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6$, a discovery that dates to Euler. But crude bounds serve their turn just as well here.) And consequently, by Levi's theorem, $\sum_n f_n(t)$ converges a.e., whence, *a fortiori*, $f_n(t) = \left[\frac{1}{n} (r_1(t) + \cdots + r_n(t)) \right]^4$ converges to zero pointwise almost everywhere in the unit interval. Or, what is very much the same thing,

$$\lim_{n \rightarrow \infty} \frac{r_1(t) + \cdots + r_n(t)}{n} = 0 \quad \text{a.e.}$$

As the Rademacher functions are related to the binary digits via $r_k(t) = 1 - 2z_k(t)$ our discovery is equivalent to the conclusion that almost all points t in

the unit interval have a density $v(t)$ and, moreover,

$$v(t) = \lim_{n \rightarrow \infty} \frac{z_1(t) + \cdots + z_n(t)}{n} = \frac{1}{2} \quad \text{a.e.}$$

THE STRONG LAW OF LARGE NUMBERS FOR COIN TOSSES *The density of 1s is equal to the density of 0s in the dyadic expansion of almost all numbers t in the unit interval. Or, in the language of probability, for almost all sequences of coin tosses (excepting only a set of sequences whose probability is identically zero), the sequence of values $\frac{1}{n}S_n$ converges to $\frac{1}{2}$.*

Thus, the binary digits 0 and 1 appear in fair and just proportion for almost all numbers in the unit interval. Equivalently, in terms of coin tossing, almost all sequences of coin tosses have heads and tails occurring in equal proportion in the long run. We express this statement compactly by saying that *the sequence of values $\frac{1}{n}S_n$ converges to $\frac{1}{2}$ with probability one, or, more compactly still, $\frac{1}{n}S_n \rightarrow \frac{1}{2}$ a.e.* This notion of convergence is much closer to the ordinary, garden variety convergence of sequences. Convergence almost everywhere guarantees the ordinary kind of convergence excepting only for an obdurate set of sequences which happily are confined within a set of probability zero.

The strong law is much more satisfying than its anaemic shadow, the weak law. It says that for almost all sequences of coin tosses, the fraction of heads not only gets arbitrarily close to one-half for a sufficiently large n , but *stays* close to one-half thereafter. This is the ultimate justification of polling mechanisms, hence its great practical importance.

The reader should ask why the apparently mysterious choice of functions f_n given by (7.1) proved to be so efficacious. A good start is to investigate what happens if we had chosen $f_n(t) = \frac{1}{n}(r_1(t) + \cdots + r_n(t))^2$ instead.

The gentle reader is doubtless aware that there is nothing sacrosanct about base 2; the decimal base 10 is in constant use in ordinary usage, while hexadecimal (base 16) and octal (base 8) bases are common in computer science. In general, any number $t \in [0, 1]$ has a unique expansion in any given base $d \geq 2$ (again adopting a terminating convention),

$$t = \sum_{k=1}^{\infty} \frac{z_k(t)}{d^k} = .z_1(t)z_2(t)\cdots \quad (\text{base } d),$$

where the d -ary digits $z_k(t)$ are integers taking values in $0, 1, \dots, d-1$ only. This is called the *d -adic expansion* of t . Let t be any number in the unit interval $[0, 1]$ and k any integer in $\{0, 1, \dots, d-1\}$. Write $S_k^{(d)}(t; n)$ for the number of occurrences of k in the first n digits of the d -adic expansion $t = .z_1(t)z_2(t)\cdots z_n(t)\cdots$. The *density of k in the d -adic expansion of t* is given by

$$v_k^{(d)}(t) = \lim_{n \rightarrow \infty} \frac{S_k^{(d)}(t; n)}{n}$$

provided the limit exists. Our definition of normality adjusts appropriately.

DEFINITION 3 A number t is *normal in base d* if each of the integers $k = 0, 1, \dots, d - 1$ has density equal to $1/d$ in the d -adic expansion of t . A number t is *normal* if it is normal in every base.

It should not be surprising now that the strong law of large numbers is not base-dependent. The following beautiful result was shown by E. Borel in 1909.

BOREL'S LAW OF NORMAL NUMBERS *Almost all numbers t in the unit interval are normal.*

More verbosely, excepting t in a set of measure zero, the d -adic expansion of t exhibits all digits from 0 to $d - 1$ in equal frequency simultaneously in all bases d . The reader is strongly encouraged to provide her own proof. It requires little more than retracing the path for base 2 and modifying it for a general base d to argue that almost all numbers are normal in base d . Thus, if $\mathbb{A}^{(d)}$ is the set consisting of the points t in the unit interval that are not normal in base d then $\lambda(\mathbb{A}^{(d)}) = 0$. The set \mathbb{A} on which t is not normal is the union over $d \geq 2$ of sets $\mathbb{A}^{(d)}$ of this form. Thus, the aberrant set \mathbb{A} of non-normal numbers has total measure no larger than a denumerable sum of the individual measures of the sets $\mathbb{A}^{(d)}$ which, however, adds up to zero as each set individually has measure zero.

8 Problems

1. *Orthogonality.* Show that for every strictly positive integer n we have

$$\int_0^1 r_{k_1}(t)r_{k_2}(t) \cdots r_{k_n}(t) dt = 0$$

whenever the indices k_1, \dots, k_n are distinct.

2. *Independent ternary digits.* Write the ternary expansion of t , $0 \leq t < 1$, in the form

$$t = \frac{\eta_1(t)}{3} + \frac{\eta_2(t)}{3^2} + \frac{\eta_3(t)}{3^3} + \dots,$$

where the *ternary digits* $\eta_k(t)$ can assume values 0, 1, and 2 only. Prove that the ternary digits $\eta_k(t)$ are independent (in the same sense that the binary digits $z_k(t)$ are independent).

3. *Generalisation of Viète's formula.* Show that

$$\frac{\sin x}{x} = \prod_{k=1}^{\infty} \frac{1 + 2 \cos \frac{2x}{3^k}}{3},$$

and generalise it. [Hint: Consider the trigonometric expansions of $\sin(\frac{x}{3} + \frac{2x}{3})$ and $\sin(\frac{x}{3} + \frac{x}{3})$.]

4. *Rademacher functions.* For each k , let $R_n(t) = r_1(t) + \dots + r_n(t)$ be the sum of the first n Rademacher functions. Evaluate $\int_0^1 R_n(t)^6 dt$.

5. *Continuation.* Apply the result of the previous problems to obtain a bound for the probability that n tosses of a fair coin result in more than $\frac{1}{2}n + \epsilon n$ heads. Compare the result with what Chebyshev's inequality gives and determine which bound is superior.

6. *Continuation.* Evaluate $I_{m,n} = \int_0^1 R_n(t)^m dt$ in terms of m and n .

7. *Numbers without density.* Exhibit a number $t = .z_1 z_2 \dots$ (base 2) in the unit interval for which the sequence of values $(z_1 + \dots + z_n)/n$ does not converge.

8. *Non-normal numbers.* Construct an example of a number t that is not normal (in any base).

9. *Walsh–Kaczmarz functions.* If n is any strictly positive integer, we may express the corresponding even positive integer $2n$ in binary notation uniquely in the form

$$2n = 2^{n_1} + 2^{n_2} + \dots + 2^{n_k}$$

where $1 \leq n_1 < n_2 < \dots < n_k$. The numbers n_1, \dots, n_k are hence unique positive integers that completely determine n . (Of course, the number of terms in the binary expansion, $k = k(n)$, depends on n .) The Walsh–Kaczmarz functions $w_n(t)$ are now defined for $0 \leq t < 1$ as follows:

$$\begin{aligned} w_0(t) &= 1, \\ w_n(t) &= r_{n_1}(t)r_{n_2}(t)\dots r_{n_k}(t) \quad (n \geq 1). \end{aligned}$$

Prove that the functions $w_n(t)$ are orthonormal, i.e.,

$$\int_0^1 w_m(t)w_n(t) dt = \begin{cases} 0 & \text{if } m \neq n, \\ 1 & \text{if } m = n. \end{cases}$$



10. Let $f(t)$ be a continuous function on the unit interval $0 \leq t \leq 1$. Prove that

$$\lim_{n \rightarrow \infty} \int_0^1 \dots \int_0^1 f\left(\frac{x_1 + \dots + x_n}{n}\right) dx_1 \dots dx_n = f\left(\frac{1}{2}\right).$$

[Hint: First prove, imitating Chebyshev's proof of the weak law of large numbers for the binomial, that the n -dimensional volume of the set defined by the inequalities

$$\left| \frac{x_1 + \dots + x_n}{n} - \frac{1}{2} \right| > \epsilon \quad (0 \leq x_i \leq 1, i = 1, \dots, n)$$

is less than $1/(12\epsilon^2 n)$. You should recall that a continuous function on a closed and bounded set is uniformly continuous: what this means is that, for every $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$ independent of x such that $|f(x+h) - f(x)| < \epsilon$ whenever $|h| < \delta$ for each x in the closed unit interval $[0, 1]$.]

11. Using the formula

$$|u| = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1 - \cos(ux)}{x^2} dx, \tag{8.1}$$

prove first that

$$\int_0^1 \left| \sum_{k=1}^n r_k(t) \right| dt = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1 - \cos(x)^n}{x^2} dx > \frac{1}{\pi} \int_{-1/\sqrt{n}}^{1/\sqrt{n}} \frac{1 - \cos(x)^n}{x^2} dx,$$

and finally that

$$\int_0^1 \left| \sum_{k=1}^n r_k(t) \right| dt > A\sqrt{n}$$

where the constant A may be chosen as

$$A = \frac{1}{\pi} \int_{-1}^1 \frac{1 - e^{-y^2/2}}{y^2} dy.$$

[Hint: Recall that $e^x = \sum_{k=0}^{\infty} x^k/k!$ and $\cos(x) = \sum_{k=0}^{\infty} (-1)^k x^{2k}/(2k)!$; these are, of course, just the familiar Taylor series formulae for the exponential and the cosine. And for the reader wondering where the formula (8.1) came from, rewrite the integral in the form

$$\frac{u^2}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\sin(ux/2)}{ux/2} \right)^2 dx,$$

and recognise the fact that the area under the sinc-squared function $\sin(\xi)^2/\xi^2$ is π . (Or, for the reader unfamiliar with Fourier analysis, peek ahead and apply Parseval's equation from Section VI.3 to the rectangular function of Example VI.2.1.)]

Problems 12–21 deal with biased coins and the properties of bent binary digits.

12. Partitions of the unit interval. Suppose $0 < p < 1$; set $q = 1 - p$. The following recursive procedure partitions the unit interval $\mathbb{I} = [0, 1)$ into a sequence of finer and finer subintervals, 2 at the first step, 4 at the second, and so on, with 2^k at step k . Begin at step 1 by setting $\mathbb{I}_0^{(1)} = [0, q)$ and $\mathbb{I}_1^{(1)} = [q, 1)$. Now suppose that at step k the 2^k intervals $\mathbb{I}_{j_1, \dots, j_k}^{(k)}$ have been specified where, in lexicographic notation, (j_1, \dots, j_k) sweeps through all 2^k binary sequences in $\{0, 1\}^k$. Each subinterval $\mathbb{I}_{j_1, \dots, j_k}^{(k)}$ (the parent) at step k engenders two subintervals (the children) at level $k + 1$, a left subinterval $\mathbb{I}_{j_1, \dots, j_k, 0}^{(k+1)}$ whose length is a fraction q of the parent and a right subinterval $\mathbb{I}_{j_1, \dots, j_k, 1}^{(k+1)}$ whose length is a fraction p of the parent. In more detail, suppose the subinterval $\mathbb{I}_{j_1, \dots, j_k}^{(k)} = [\alpha, \beta)$ has left endpoint α and length $\beta - \alpha$. Then the left subinterval engendered by it is specified by $\mathbb{I}_{j_1, \dots, j_k, 0}^{(k+1)} = [\alpha, \alpha + q(\beta - \alpha))$ with the corresponding right subinterval given by $\mathbb{I}_{j_1, \dots, j_k, 1}^{(k+1)} = [\alpha + q(\beta - \alpha), \beta)$. In this fashion, the 2^{k+1} subintervals $\mathbb{I}_{j_1, \dots, j_k, j_{k+1}}^{(k+1)}$ at step $k + 1$ are engendered. Show that $\mathbb{I}_{j_1, \dots, j_k}^{(k)}$ has length $p^{j_1 + \dots + j_k} q^{k-j_1-\dots-j_k}$.

13. Continuation, bent binary digits. For each $k \geq 1$, define the bent binary digit $z_k(t; p)$ by setting $z_k(t; p) = 0$ if $t \in \mathbb{I}_{j_1, \dots, j_{k-1}, 0}^{(k)}$ for some specification of bits j_1, \dots, j_{k-1} , and $z_k(t; p) = 1$ if, instead, $t \in \mathbb{I}_{j_1, \dots, j_{k-1}, 1}^{(k)}$. These functions coincide with the ordinary binary digits $z_k(t)$ if $p = 1/2$. Show that the bent binary digits $z_k(t; p)$ ($k \geq 1$) are independent for each value of p .

14. *Non-symmetric binomial distribution.* With $z_k(t; p)$ the bent binary digits of the previous problem, prove that the measure of the set of points t on which $z_1(t; p) + \dots + z_n(t; p) = k$ is given by

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikx} (pe^{ix} + q)^n dx$$

and explicitly evaluate this expression for each integer k .

15. *Continuation, unfair coins.* Explain how the bent binary digits $z_k(t; p)$ ($k \geq 1$) can be used to construct a model for independent tosses of an unfair coin whose success probability is p and failure probability is $q = 1 - p$.

16. *Bernstein polynomials.* Suppose $f(t)$ is a continuous function on the unit interval. Show that

$$\int_0^1 f\left(\frac{z_1(t; x) + \dots + z_n(t; x)}{n}\right) dt = \sum_{k=0}^n f\left(\frac{k}{n}\right) x^k (1-x)^{n-k}.$$

For each n , the expression $B_n(x)$ given by the sum on the right defines a function on the unit interval $0 \leq x \leq 1$ [with the natural convention $0^0 = 1$ giving the values $B_n(0) = f(0)$ and $B_n(1) = f(1)$.] The functions $B_n(x)$ are the famous Bernstein polynomials.

17. *Weak law of large numbers.* Let S_n be the number of successes in n tosses of a bent coin whose success probability is p . By using Chebyshev's idea, show that $\frac{1}{n} S_n$ converges in probability to p . In other words, $P\left\{ \left| \frac{1}{n} S_n - p \right| \geq \epsilon \right\} \rightarrow 0$ as $n \rightarrow \infty$ for every choice of $\epsilon > 0$.

18. *Strong law of large numbers.* Prove that $\frac{1}{n} S_n$ converges almost everywhere to p . This substantially strengthens the result of Problem 17.

 **19.** *Weierstrass's approximation theorem.* Use the results of Problem 17 to estimate the measure of the set on which

$$\left| \frac{z_1(t; x) + \dots + z_n(t; x)}{n} - x \right| > \epsilon$$

and prove thence that $\lim_{n \rightarrow \infty} B_n(x) = f(x)$ uniformly in $0 \leq x \leq 1$. This is Bernstein's original proof of the celebrated theorem of Weierstrass on the approximation of continuous functions by polynomials.

 **20.** *Continuation, Lipschitz functions.* We say that a real-valued function f is *Lipschitz* if there exists a constant M such that $|f(x) - f(y)| \leq M|x - y|$ for all choices of x and y . If $f: [0, 1] \rightarrow \mathbb{R}$ is Lipschitz with constant M , show that $|f(x) - B_n(x)| \leq M/(2\sqrt{n})$.

 **21.** *Continuation, tightness of the estimate.* Let $f(x) = |x - 1/2|$ on the unit interval $0 \leq x \leq 1$. Use the result of Problem 11 to estimate $|f(1/2) - B_n(1/2)|$ from below and thus show that the order $1/\sqrt{n}$ error estimate of the previous problem is tight.

The Normal Law

The normal law made its appearance as early as 1733 in the work of Abraham de Moivre as an approximation to probabilities involving the number of successes in a succession of tosses of a fair coin. De Moivre's result was gradually generalised and extended by degrees until it became clear that the result was of quite extraordinary generality and power. The normal law, also called the *central limit theorem*, stands now in the rôle of an *éminence grise* in probability theory—there is scarce a branch of science and engineering that has not been impacted by it.

c 1–5

§ 6–9

I have chosen to provide a proof of de Moivre's theorem in this chapter (though not using his methods) both to further illustrate the utility of the correspondence between coin tosses and numbers that we explored in the previous chapter as well as to give the reader early access to the original version of the theorem that is still of frequent use in applications. The mathematics is a little harder, to be sure, than what we have encountered hitherto, but the reader may feel that the price is well worth the paying to get an early exposure to a fundamental pillar of probability—and in a context of particular importance. The reader anxious to see the theorem in action will encounter two topical applications in Sections 8 and 9. We shall come back to the subject armed with more background in Chapter XX where we will prove several general versions of this important theorem and see applications in divers areas.

Before proceeding to explore the fine structure of the distribution of successes in a sequence of coin tosses it will be useful to digress briefly to collect some basic facts from Fourier theory. The reader who is familiar with Fourier methods should, after reading the next section, skip on to Section 4, dipping back at need to refresh her memory.

1 One curve to rule them all

We are grown to casual familiarity with references to the bell curve in the popular press. (And, if a student has taken a first course in statistics, she may be

forgiven for leaving with the feeling that the entire discipline is an extended excursus on the curve.) Formally speaking, the bell curve, or, to give it its proper name, the *standard normal density* (also called the *Gaussian density*) is the real-valued function $\phi(x)$ of a real variable x defined by

$$\phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}. \quad (1.1)$$

Its graph, shown in Figure 1, has the familiar bell shape, whence its popular

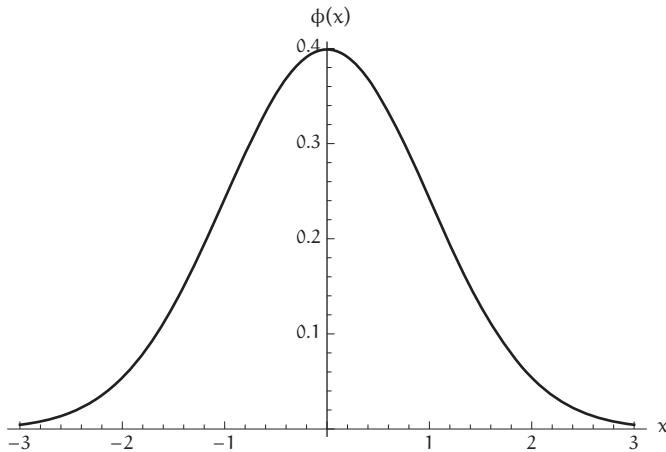


Figure 1: The standard normal density.

monicker. This curve has a peculiar and central importance in probability and, to save on future repetition, it will be useful to collect the basic facts regarding it.

It is clear that ϕ is an even function, strictly positive everywhere on the real line; and, in view of the rapid extinction of the exponential, it is evident that it dies monotonically and rapidly away from the origin on both sides. It follows by the integral test that the area under the curve is positive and bounded. A simple trick shows indeed that its area is properly normalised.

LEMMA 1 *The area under the curve of the standard normal density is unit; formally,*

$$\int_{-\infty}^{\infty} \phi(x) dx = 1.$$

PROOF: As ϕ is positive it is clear that the integral on the left is positive. By considering its *square*, we obtain

$$\begin{aligned} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx \right)^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-x^2/2} dx \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \frac{1}{2\pi} \iint_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dy dx = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta = 1, \end{aligned}$$

where the rectangular-to-polar transformation $x = r \cos \theta$ and $y = r \sin \theta$ in the penultimate step is both natural and efficacious. It follows that $\int_{-\infty}^{\infty} \phi(x) dx = 1$ and the awkward-looking normalisation by $(2\pi)^{-1/2}$ is validated. ►

For each integer $j \geq 0$, define the j th moment of the normal by

$$M_j = \int_{-\infty}^{\infty} x^j \phi(x) dx.$$

As, for each j , we can find a constant A_j such that $|x|^j \leq A_j e^{|x|}$, the comparison test shows that the integral is convergent for each j .

LEMMA 2 If j is odd then $M_j = 0$; if j is even then $M_j = j! / (2^{j/2} (j/2)!)$. In particular, $M_0 = 1$, $M_1 = 0$, and $M_2 = 1$.

PROOF: If j is odd then the monomial x^j is an odd function and as $\phi(x)$ is an even function, the product $x^j \phi(x)$ is odd, hence has zero area.

If $j = 2v$ is even then a simple integration by parts sets up an induction. For each $v \geq 1$, we have

$$\begin{aligned} M_{2v} &= \int_{-\infty}^{\infty} x^{2v} \phi(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2v-1} d(-e^{-x^2/2}) \\ &= \frac{-x^{2v-1} e^{-x^2/2}}{\sqrt{2\pi}} \Big|_{-\infty}^{\infty} + \frac{2v-1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2v-2} e^{-x^2/2} dx = (2v-1) M_{2v-2}. \end{aligned}$$

As $M_0 = 1$ by the previous lemma, it follows readily by induction that

$$M_{2v} = (2v-1)(2v-3) \cdots 5 \cdot 3 \cdot 1 \cdot M_0.$$

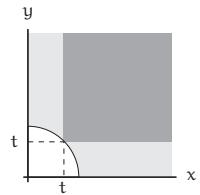
On the right we have all the odd terms in the product for $(2v)!$. If we insert the missing even terms and compensate by dividing by them as well, then in the numerator we obtain $(2v)!$ while, by factoring out 2 from each of the even terms, in the denominator we obtain $2^v v!$ as was to be shown. ►

As we have remarked earlier, the normal curve exhibits a very rapid decay away from the origin. While there are several ways to verify this analytically, the following result has the great virtue of simplicity.

LEMMA 3 The inequality $\int_t^{\infty} \phi(x) dx \leq \frac{1}{2} e^{-t^2/2}$ holds for every $t \geq 0$.

PROOF: A geometric trick helps reduce the computation to that of an elementary integral. By squaring the integral again, we observe that

$$\left(\int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \right)^2 = \iint_{\mathbb{A}} \frac{e^{-(x^2+y^2)/2}}{2\pi} dy dx$$



where \mathbb{A} is the rectangular quadrant $\{(x, y) : x > t, y > t\}$ shown heavily shaded in Figure 2. It is clear that \mathbb{A} is contained in the region $\mathbb{B} = \{(x, y) : x^2 + y^2 > 2t^2\}$ which includes the additional region around \mathbb{A} shown lightly shaded. It follows via a change to polar coordinates that

$$\left(\int_t^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx \right)^2 \leq \iint_{\mathbb{B}} \frac{e^{-(x^2+y^2)/2}}{2\pi} dy dx = \int_0^{\pi/2} \int_{\sqrt{2}t}^\infty \frac{e^{-r^2/2}}{2\pi} r dr d\theta = \frac{e^{-t^2}}{4}.$$

Taking square-roots of both sides completes the proof. ▶

The simple nature of the bound makes it very useful in practice and, as we shall see in Section X.1, the estimate is quite remarkably good.

2 A little Fourier theory I

Suppose $f(x)$ is a complex-valued function of a real variable x . For the purposes of the next few sections, we say that f is *integrable* if it is Riemann integrable over every bounded interval $[a, b]$ and $\int_{-\infty}^{\infty} |f(x)| dx$ converges. If f is integrable, the *Fourier transform* of f , denoted \hat{f} , is the complex-valued function of a real variable defined formally by¹

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{i\xi x} dx.$$

The transform was introduced by Jean Baptiste Joseph Fourier in 1811 and in its lasting influence over the developments of a couple of centuries can lay some claim to being the most influential equation of them all.

It is of key analytical importance that the Fourier transforms of a variety of simple operators have simple forms. We consider some of the more common operations.

¹It is usual in probability theory to define the transform with a positive sign, $+i\xi x$, in the exponent as given here. In Fourier analysis, however, it is more usual to define the transform with a negative sign, $-i\xi x$, in the exponent. It does not make much matter which definition we choose and we may as well hew to probabilistic convention.

Translation: For any real x_0 , the translation map \mathcal{T}_{x_0} takes each function f into the function $\mathcal{T}_{x_0}f(x) = f(x - x_0)$. The map \mathcal{T}_{x_0} effects a shift of origin to the point x_0 .

Scale: For any $\sigma > 0$, the scale map \mathcal{S}_σ takes each f into the function $\mathcal{S}_\sigma f(x) = \frac{1}{\sigma} f(\frac{x}{\sigma})$. The map \mathcal{S}_σ preserves area and effects a compression (if $0 < \sigma < 1$) or dilation (if $\sigma > 1$) of the axis.

Convolution: The convolution operator \star takes each pair of integrable and bounded functions f and g into the function $f \star g$ defined by

$$f \star g(x) = \int_{-\infty}^{\infty} f(t)g(x-t) dt = \int_{-\infty}^{\infty} f(x-t)g(t) dt,$$

the two integral forms obtained one from the other by a simple change of variable inside the integral. We call $f \star g$ the convolution of f and g .

Differentiation: The differentiation operator \mathcal{D} maps each differentiable f into its derivative, $\mathcal{D}f(x) = f'(x) = df(x)/dx$.

THEOREM 1 (PROPERTIES OF THE FOURIER TRANSFORM)

1. *Linearity:* If f and g are integrable and c a complex constant then $\widehat{f+g}(\xi) = \widehat{f}(\xi) + \widehat{g}(\xi)$ and $\widehat{c \cdot f}(\xi) = c \cdot \widehat{f}(\xi)$.
2. *Translation:* If f is integrable and x_0 is real then $\widehat{\mathcal{T}_{x_0}f}(\xi) = e^{ix_0\xi} \widehat{f}(\xi)$.
3. *Scale:* If f is integrable and $\sigma > 0$ then $\widehat{\mathcal{S}_\sigma f}(\xi) = \widehat{f}(\sigma\xi)$.
4. *Convolution:* If f and g are integrable and bounded then $\widehat{f \star g}(\xi) = \widehat{f}(\xi)\widehat{g}(\xi)$.
5. *Differentiation I:* If f is continuously differentiable, f and $\mathcal{D}f$ are integrable, and $f(x) \rightarrow 0$ as $|x| \rightarrow \infty$, then $\widehat{\mathcal{D}f}(\xi) = -i\xi \widehat{f}(\xi)$.
6. *Differentiation II:* If $f(x)$ and $xf(x)$ are both integrable then \widehat{f} is differentiable and, moreover, its derivative $\widehat{\mathcal{D}f}(\xi)$ is the Fourier transform of $ixf(x)$.

PROOF: Linearity follows easily as Riemann integration is a linear operation and the translation and scale properties are easy to verify by a change of variable inside the Fourier integral. To verify the convolution property, we check first that

$$\int_{-\infty}^{\infty} |f(x-t)| \cdot |g(t)| dx = |g(t)| \int_{-\infty}^{\infty} |f(x-t)| dx$$

is bounded and as g , hence also $|g|$, is integrable, it follows that

$$\int_{-\infty}^{\infty} |g(t)| \int_{-\infty}^{\infty} |f(x-t)| dx dt$$

converges. Consequently,

$$\begin{aligned}\widehat{f \star g}(\xi) &= \int_{-\infty}^{\infty} f \star g(x) e^{i \xi x} dx = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x-t) g(t) dt \right) e^{i \xi x} dx \\ &= \int_{-\infty}^{\infty} g(t) \left(\int_{-\infty}^{\infty} f(x-t) e^{i \xi (x-t)} dx \right) e^{i \xi t} dt\end{aligned}$$

as $f \star g$ is bounded and integrable and it is hence legitimate to change the order of integration. Via the change of variable $y = x - t$ in the inner integral on the right we obtain

$$\widehat{f \star g}(\xi) = \int_{-\infty}^{\infty} g(t) e^{i \xi t} dt \int_{-\infty}^{\infty} f(y) e^{i \xi y} dy = \widehat{f}(\xi) \widehat{g}(\xi)$$

as was to be shown.

Finally, to verify the differentiation property, an integration by parts shows that

$$\widehat{\mathcal{D}f}(\xi) = \int_{-\infty}^{\infty} f'(x) e^{i \xi x} dx = f(x) e^{i \xi x} \Big|_{-\infty}^{\infty} - i \xi \int_{-\infty}^{\infty} f(x) e^{i \xi x} dx = -i \xi \widehat{f}(\xi)$$

as f vanishes at infinity. To verify the dual statement, we begin with the Fourier transform

$$\widehat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{i \xi x} dx.$$

Differentiating both sides with respect to ξ , we obtain

$$\mathcal{D}\widehat{f}(\xi) = \frac{d}{d\xi} \int_{-\infty}^{\infty} f(x) e^{i \xi x} dx \stackrel{(\dagger)}{=} \int_{-\infty}^{\infty} f(x) \frac{\partial}{\partial \xi} e^{i \xi x} dx = i \int_{-\infty}^{\infty} x f(x) e^{i \xi x} dx$$

which is what was to be established. The only step that needs justification is the “obvious” interchange of differentiation and integration in the step marked (\dagger) . The reader may be willing to accept it but, if she is feeling particularly picky, she may work through the proof given below or, even better, provide a proof for herself. ▶

LEMMA *If $f(x)$ and $xf(x)$ are both integrable then*

$$\frac{d}{d\xi} \int_{-\infty}^{\infty} f(x) e^{i \xi x} dx = \int_{-\infty}^{\infty} f(x) \frac{\partial}{\partial \xi} e^{i \xi x} dx.$$

PROOF: Here is a first principles argument. For any fixed, small $\zeta \neq 0$,

$$\begin{aligned}\frac{\widehat{f}(\xi + \zeta) - \widehat{f}(\xi)}{\zeta} &= \frac{1}{\zeta} \int_{-\infty}^{\infty} f(x) e^{i(\xi+\zeta)x} dx - \frac{1}{\zeta} \int_{-\infty}^{\infty} f(x) e^{i \xi x} dx \\ &= \int_{-\infty}^{\infty} f(x) \left(\frac{e^{i(\xi+\zeta)x} - e^{i \xi x}}{\zeta} \right) dx = \int_{-\infty}^{\infty} f(x) e^{i \xi x} \left(\frac{e^{i \zeta x} - 1}{\zeta} \right) dx.\end{aligned}$$

As $e^{i\zeta x}$ has unit modulus, it follows that

$$\begin{aligned} \left| \frac{\widehat{f}(\xi + \zeta) - \widehat{f}(\xi)}{\zeta} - \int_{-\infty}^{\infty} ix f(x) e^{i\zeta x} dx \right| &= \left| \int_{-\infty}^{\infty} f(x) e^{i\zeta x} \left(\frac{e^{i\zeta x} - 1}{\zeta} - ix \right) dx \right| \\ &\leq \int_{-\infty}^{\infty} |f(x)| \cdot \left| \frac{e^{i\zeta x} - 1}{\zeta} - ix \right| dx. \end{aligned} \quad (2.1)$$

The idea now is to show that the integrand on the right is small in a neighbourhood of the origin, the integral tails contributing little. Taylor's formula provides the key.

Fix any x and consider the exponential function $g(\zeta) = e^{i\zeta x}$ viewed as a function of ζ (with x as a fixed parameter). As $g'(\zeta) = ix e^{i\zeta x}$ and $g''(\zeta) = -x^2 e^{i\zeta x}$, Taylor's formula truncated first at the linear term and then at the quadratic term (or, equivalently, two applications of the mean value theorem) shows that we may write

$$e^{i\zeta x} = 1 + ix\zeta e^{i\zeta_0 x} = 1 + ix\zeta - \frac{1}{2}x^2\zeta^2 e^{i\zeta_1 x}$$

for some ζ_0 and ζ_1 lying between 0 and ζ . We may now estimate the second term in the integrand on the right in (2.1) in two different ways. Suppose $\zeta \neq 0$. Then

$$\left| \frac{e^{i\zeta x} - 1}{\zeta} - ix \right| = \begin{cases} |x| \cdot |e^{i\zeta_0 x} - 1| \leq 2|x|, \\ \left| -\frac{1}{2}x^2\zeta e^{i\zeta_1 x} \right| = \frac{1}{2}x^2|\zeta|. \end{cases} \quad (2.2)$$

The stage is set for a divide-and-conquer argument. The bound originating from the linear Taylor approximation in the first line in (2.2) is a little generous to be sure but suffices to estimate the contribution of the integral tails on the right in (2.1). As $xf(x)$ is integrable, $\int_{-\infty}^{\infty} |xf(x)| dx$ converges and, *a fortiori*, the integral tails vanish. In particular, for every $\epsilon > 0$, there exists $T = T(\epsilon)$ such that $\int_{|x| \geq T} |xf(x)| dx < \epsilon$. It follows that

$$\int_{|x| \geq T} |f(x)| \cdot \left| \frac{e^{i\zeta x} - 1}{\zeta} - ix \right| dx \leq 2 \int_{|x| \geq T} |xf(x)| dx < 2\epsilon.$$

We obtain a more refined estimate of the integrand in the neighbourhood $|x| < T$ of the origin via the quadratic Taylor approximation in the second line in (2.2). As $f(x)$ is integrable, $\int_{-\infty}^{\infty} |f(x)| dx$ converges and is equal to a finite value, say, M . Accordingly,

$$\int_{-T}^T |f(x)| \cdot \left| \frac{e^{i\zeta x} - 1}{\zeta} - ix \right| dx = \frac{|\zeta|}{2} \int_{-T}^T x^2 |f(x)| dx \leq \frac{|\zeta|T^2}{2} \int_{-T}^T |f(x)| dx \leq \frac{|\zeta|M T^2}{2}.$$

The right-hand side may be made as small as desired by taking ζ small; in particular, if $0 < |\zeta| < 2\epsilon/M T^2$ then the bound on the right is $< \epsilon$. Summing the contributions from the regions $|x| < T$ and $|x| \geq T$ to the right-side of (2.1), we obtain

$$\left| \frac{\widehat{f}(\xi + \zeta) - \widehat{f}(\xi)}{\zeta} - \int_{-\infty}^{\infty} ix f(x) e^{i\zeta x} dx \right| < 3\epsilon$$

for all ζ sufficiently small in absolute value. As $\epsilon > 0$ may be chosen arbitrarily small, it follows by passing to the limit as $\zeta \rightarrow 0$ that

$$\frac{d\widehat{f}(\xi)}{d\xi} = \lim_{\zeta \rightarrow 0} \frac{\widehat{f}(\xi + \zeta) - \widehat{f}(\xi)}{\zeta} = \int_{-\infty}^{\infty} ix f(x) e^{i\zeta x} dx$$

as was to be established. ►

It is clear that a more general theorem concerning differentiating under the integral sign lurks under the analysis, but we won't pause to pull it out.

The examples given below have the virtue of both being useful and illustrating the properties in action.

EXAMPLES: 1) *The rectangular function.* We define the unit rectangular function by

$$\text{rect}(x) = \begin{cases} 1 & \text{if } |x| < 1/2, \\ 0 & \text{if } |x| \geq 1/2. \end{cases}$$

Determining its Fourier transform is just a matter of integrating the exponential and we obtain

$$\widehat{\text{rect}}(\xi) = \frac{e^{i\xi/2} - e^{-i\xi/2}}{i\xi} = \frac{\sin(\xi/2)}{\xi/2}$$

where we interpret the expressions on the right to be equal to 1 when $\xi = 0$. (The reader can easily verify by l'Hôpital's rule that this does no violence to the spirit of the expression on the right.) We can compact notation by defining the *sinc function* by

$$\text{sinc}(\xi) = \frac{\sin(\xi)}{\xi},$$

with the convention that the right-hand side is to be interpreted as equal to 1 when $\xi = 0$. Then, $\widehat{\text{rect}}(\xi) = \text{sinc}(\xi/2)$.

2) *The triangular function.* The unit triangular function $\Delta(x) = \max\{0, 1 - |x|\}$ has a triangular graph with base $-1 < x < 1$ and apex of height 1 at the origin. While its transform may be readily obtained by an integration by parts, a slicker path is accorded by the realisation that $\Delta(x) = \text{rect} * \text{rect}(x)$ is the convolution of the unit rectangular function with itself and, in view of the convolution property, we may directly write down $\widehat{\Delta}(\xi) = \widehat{\text{rect}}(\xi)^2 = \text{sinc}(\xi/2)^2$.

3) *The trapezoidal function.* Consider the function obtained by smoothing out the corners of the unit rectangle function to form a trapezoidal graph as shown in Figure 3. While its Fourier transform can be computed by a direct, if slightly tedious, integration by splitting up the region of integration, it is simpler to express the trapezoidal function as the difference of two triangular functions as shown by the dotted lines in the graph. For each non-zero τ , we define

$$\text{trap}_\tau(x) = \frac{1/2 + \tau}{\tau} \Delta\left(\frac{x}{1/2 + \tau}\right) - \frac{1/2}{\tau} \Delta\left(\frac{x}{1/2}\right),$$

where, in the notation of Example 2, $\Delta(x)$ represents the unit triangular function with a base of two units and a height of one unit. As the reader may quickly verify, when $\tau = +\epsilon$ is positive we obtain the trapezoidal graph of height 1,

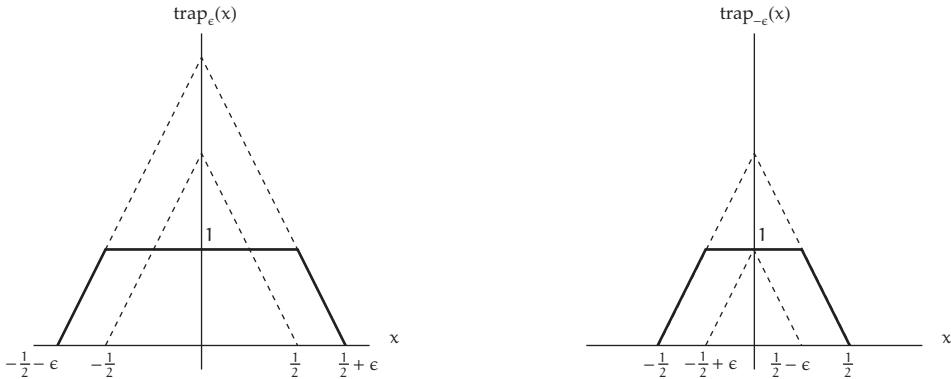


Figure 3: Two versions of the trapezoidal function for a positive ϵ .

floor $1 + 2\epsilon$, and roof 1 as shown on the left, while when $\tau = -\epsilon$ is negative we obtain the trapezoidal graph of height 1, floor 1, and roof $1 - 2\epsilon$ as shown on the right. The scaling and linearity properties of the transform immediately allow us to write down the Fourier transform. As $\widehat{\Delta}(\xi) = \text{sinc}(\xi/2)^2$, we obtain

$$\begin{aligned}\widehat{\text{trap}}_\tau(\xi) &= \frac{(1/2 + \tau)^2}{\tau} \widehat{\Delta}\left[\left(\frac{1}{2} + \tau\right)\xi\right] - \frac{(1/2)^2}{\tau} \widehat{\Delta}\left[\left(\frac{1}{2}\right)\xi\right] \\ &= \frac{(1 + 2\tau)^2}{4\tau} \text{sinc}\left(\frac{(1 + 2\tau)\xi}{4}\right)^2 - \frac{1}{4\tau} \text{sinc}\left(\frac{\xi}{4}\right)^2 = \frac{4}{\tau\xi^2} \sin\left(\frac{\tau\xi}{2}\right) \sin\left(\frac{(1 + \tau)\xi}{2}\right),\end{aligned}$$

the simplification in the final step following by the elementary trigonometric identities $2\sin(A)^2 = 1 - \cos(2A)$ and $2\sin(A)\sin(B) = \cos(A - B) - \cos(A + B)$.

4) *The standard normal density.* The expression

$$\widehat{\phi}(\xi) = \int_{-\infty}^{\infty} \phi(x) e^{i\xi x} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi x - x^2/2} dx \quad (2.3)$$

is called the *characteristic function of the normal* (or, simply, the *Gaussian characteristic function*) and represents the Fourier transform of the standard normal density $\phi(x)$. The usual approach to computing $\widehat{\phi}$ is by complex variable methods by integration over a suitable contour in the complex plane. The derivative property provides an alternative tack that the reader may find she prefers.

The function ϕ is as smooth as may be and has derivatives of all orders. In particular, taking derivatives of both sides of (1.1), we see that ϕ satisfies the differential equation

$$\mathcal{D}\phi(x) = -x\phi(x).$$

It is clear that both ϕ and $\mathcal{D}\phi$ are continuous, integrable, and vanish at infinity (because of the rapid extinction of the exponential away from the origin).

Taking the Fourier transform of both sides of the given differential equation, on the left we obtain $-i\xi\hat{\phi}(\xi)$ while on the right we obtain $i\mathcal{D}\hat{\phi}(\xi)$ in view of the derivative property and its dual formulation. We hence obtain

$$\mathcal{D}\hat{\phi}(\xi) = -\xi\hat{\phi}(\xi)$$

and, remarkably, the Fourier transform $\hat{\phi}$ also satisfies the same differential equation as ϕ . The general solution is given by $\hat{\phi}(\xi) = \hat{\phi}(0)e^{-\xi^2/2}$ and it only remains to determine $\hat{\phi}(0)$. But, substituting $\xi = 0$ in (2.3), by Lemma 1.1 it is clear that $\hat{\phi}(0) = \int_{-\infty}^{\infty} \phi(x) dx = 1$. It follows that

$$\hat{\phi}(\xi) = e^{-\xi^2/2} = \sqrt{2\pi} \phi(\xi).$$

The scale property of the transform shows perforce that, for $\sigma > 0$, the function $S_\sigma\phi(x) = \frac{1}{\sigma}\phi(\frac{x}{\sigma})$ has Fourier transform $\widehat{S_\sigma\phi}(\xi) = \hat{\phi}(\sigma\xi) = e^{-\sigma^2\xi^2/2}$. ▶

If we write $\mathcal{F}: \phi \mapsto \hat{\phi}$ for the Fourier transform operator, we may restate our findings of the last example in the form $\mathcal{F}\phi(\xi) = \sqrt{2\pi}\phi(\xi)$ or, in other words, *the normal (or Gaussian) density is an eigenfunction of the Fourier operator*. An important consequence of this is the observation that the Fourier transform relation (2.3) has the associated inverse

$$\phi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{\phi}(\xi) e^{-i\xi x} d\xi, \quad (2.3')$$

a fact which is easy to see as, by interchanging the rôles of ξ and x , the left-hand side is equal to $\hat{\phi}(x)/\sqrt{2\pi}$, while the integrand $\hat{\phi}(\xi)$ on the right may be replaced by $\sqrt{2\pi}\phi(\xi)$. As ϕ is an even function, the change of variable $\xi \leftarrow -\xi$ recovers (2.3). The Gaussian inversion formula sets the stage for inversion in a general setting.

3 A little Fourier theory II

If f is integrable then its Fourier transform \hat{f} is well-defined. Right from the genesis of the transform in 1811, a new question arose.

QUESTION A *Does \hat{f} determine f , at least when f is continuous?*

Of course, the reader may wonder what it means for \hat{f} to determine f . From a computational perspective one may wish for an explicit inversion formula to be able to invert the Fourier operation and recover f .

QUESTION B *If f is continuous, does the inversion formula $\frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-i\xi x} d\xi$ converge to $f(x)$?*

(The answers are “Yes” and, somewhat disquietingly, “Not always”, respectively.) The search for answers to these questions occupied some of the most distinguished mathematicians in history for the better part of a century and a half and resulted in an upheaval in the very foundations of mathematics. The search culminated in a remarkable paper of Lennart Carleson in 1966 which provided a comprehensive answer. As T. W. Körner has remarked, very few questions have managed to occupy even a small part of humanity for 150 years. This is one of the few.

We won’t attempt to do full justice to this glorious tapestry but will be satisfied with an inversion formula when f is particularly well behaved. We begin with a simple observation. As usual, if z is complex we write z^* for its complex conjugate.

LEMMA Suppose f , g , \hat{f} , and \hat{g} are all continuous and integrable. Then

$$\int_{-\infty}^{\infty} \hat{f}(\xi) g(\xi)^* e^{-i\xi x} d\xi = f * \hat{g}^*(x). \quad (3.1)$$

PROOF: In view of the given regularity conditions, $\iint_{-\infty}^{\infty} |f(y)| \cdot |g(s)| ds dy$ converges. Writing out the Fourier integral for \hat{f} on the left, it is hence permissible to interchange the order of integration to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(y) e^{i\xi y} dy \right) g(\xi)^* e^{-i\xi x} d\xi &= \int_{-\infty}^{\infty} f(y) \left(\int_{-\infty}^{\infty} g(\xi)^* e^{i\xi(y-x)} d\xi \right) dy \\ &= \int_{-\infty}^{\infty} f(y) \left(\int_{-\infty}^{\infty} g(\xi) e^{i\xi(x-y)} d\xi \right)^* dy = \int_{-\infty}^{\infty} f(y) \hat{g}(x-y)^* dy. \end{aligned}$$

The expression on the right is just the convolution of f with \hat{g}^* . ▶

If for \hat{g} we select a function of unit area concentrated near the origin, then the convolution on the right of the lemma will return approximately $f(x)$; the function \hat{g} will serve in the rôle of an approximate identity for convolution (or, in the language of the physicist, an approximation to a “delta function”). But if \hat{g} is concentrated near zero then g will be approximately constant over a wide range and the integral on the left of the lemma will take on the form of the Fourier inversion formula. We proceed to make this intuition precise.

Let $\{g_\sigma, \sigma > 0\}$ be a family of positive, integrable functions of a real variable normalised so that $\int_{-\infty}^{\infty} g_\sigma(x) dx = 1$ for each σ .

DEFINITION We say that $\{g_\sigma, \sigma > 0\}$ is *impulsive* if, for any $\delta > 0$, as $\sigma \rightarrow 0$, (i) $g_\sigma(x) \rightarrow 0$ uniformly for $|x| \geq \delta$, and (ii) $\int_{|x| \geq \delta} g_\sigma(x) dx \rightarrow 0$.

Impulsive families are concentrated at the origin and behave like “delta functions”. The intuition is formalised in

THE SPOTLIGHT THEOREM Suppose $\{g_\sigma, \sigma > 0\}$ is an impulsive family of functions. If $f: \mathbb{R} \rightarrow \mathbb{C}$ is integrable and continuous at the origin then

$$\int_{-\infty}^{\infty} f(x)g_\sigma(x) dx \rightarrow f(0) \quad \text{as } \sigma \rightarrow 0.$$

If f is continuous (everywhere) then $(f * g_\sigma)(x) \rightarrow f(x)$ for all x , the convergence being uniform if f is uniformly continuous.

Spelled out, if f is continuous and integrable then

$$\int_{-\infty}^{\infty} f(x-t)g_\sigma(t) dt \rightarrow f(x) \quad \text{as } \sigma \rightarrow 0,$$

and the reason for the name I have given the theorem becomes apparent—for any given x it is as if the impulsive function g_σ shines a spotlight on the value of f at x . The result is also called the *sifting theorem*, the analogy being that g_σ sifts through the values of f .

Now that we have understood what is going on it is easy to craft a proof; it follows the usual pattern of divide-and-conquer and the reader comfortable with the mechanism can move on.

PROOF: Fix any $\epsilon > 0$. Suppose f is integrable and continuous at the origin. Then $|f(x) - f(0)| \leq \epsilon$ for all x in a neighbourhood of the origin $-\delta \leq x \leq \delta$ for some choice of positive $\delta = \delta(\epsilon)$. Pick σ sufficiently small so that $0 \leq g_\sigma(x) \leq \epsilon$ for all $|x| > \delta$ and $\int_{|x|>\delta} g_\sigma(x) dx \leq \epsilon$. Then

$$\left| \int_{-\infty}^{\infty} f(x)g_\sigma(x) dx - f(0) \right| = \left| \int_{-\infty}^{\infty} (f(x) - f(0))g_\sigma(x) dx \right| \leq \int_{-\infty}^{\infty} |f(x) - f(0)|g_\sigma(x) dx.$$

It is now natural to partition the range of integration on the right into the regions $|x| \leq \delta$ and $|x| > \delta$. In the neighbourhood $[-\delta, \delta]$ around the origin, we have

$$\int_{-\delta}^{\delta} |f(x) - f(0)|g_\sigma(x) dx \leq \epsilon \int_{-\delta}^{\delta} g_\sigma(x) dx \leq \epsilon \int_{-\infty}^{\infty} g_\sigma(x) dx = \epsilon,$$

while away from the origin we may estimate the contribution to the integral by

$$\begin{aligned} \int_{|x|>\delta} |f(x) - f(0)|g_\sigma(x) dx &\leq \int_{|x|>\delta} |f(x)|g_\sigma(x) dx + |f(0)| \int_{|x|>\delta} g_\sigma(x) dx \\ &\leq \epsilon \int_{|x|>\delta} |f(x)| dx + \epsilon |f(0)| \leq \epsilon M \end{aligned}$$

where $M = |f(0)| + \int_{-\infty}^{\infty} |f(x)| dx$ is finite. It follows that

$$\left| \int_{-\infty}^{\infty} f(x)g_\sigma(x) dx - f(0) \right| \leq \epsilon(M + 1)$$

for all sufficiently small σ . As the tiny ϵ may be chosen arbitrarily small we conclude that $\int_{-\infty}^{\infty} f(x)g_{\sigma}(x) dx \rightarrow f(0)$ as $\sigma \rightarrow 0$.

If, additionally, f is continuous everywhere then, for each fixed x , we may set $f_x(y) = f(x - y)$ whence f_x is integrable and continuous at the origin. Consequently, with the change of variable $y = x - t$ inside the integral, we have

$$\begin{aligned} (f * g_{\sigma})(x) &= \int_{-\infty}^{\infty} f(t)g_{\sigma}(x-t) dt \\ &= \int_{-\infty}^{\infty} f(x-y)g_{\sigma}(y) dy = \int_{-\infty}^{\infty} f_x(y)g_{\sigma}(y) dy \rightarrow f_x(0) = f(x) \end{aligned}$$

as $\sigma \rightarrow 0$. It is easy to see that the convergence is uniform if f is uniformly continuous as in that case we may select the neighbourhood size δ as a function of ϵ alone, independently of x . ▶

If g is integrable with unit area and vanishes suitably quickly at infinity, then a coordinate scale via $g_{\sigma} = S_{\sigma}g$ provides a simple way of constructing impulsive families. In particular, starting with the normal density $\phi(x)$, it is easy to see that the coordinate-scaled family of functions $S_{\sigma}\phi(x) = \frac{1}{\sigma}\phi(\frac{x}{\sigma})$ is impulsive. If on the right in (3.1) we now identify the function \hat{g} with the impulsive function $S_{\sigma}\phi$ then, via the Gaussian inversion formula (2.3'), the corresponding function g is given by $g(t) = \frac{1}{\sqrt{2\pi}}\phi(\sigma t)$. Thus, (3.1) specialised to the Gaussian impulsive family yields the identity

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\xi)\phi(\sigma\xi)e^{-i\xi x} d\xi = f * S_{\sigma}\phi(x). \quad (3.2)$$

THE SIMPLEST FOURIER INVERSION THEOREM *If f and \hat{f} are both continuous and integrable then*

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi)e^{-i\xi x} d\xi = f(x),$$

the integral being uniformly absolutely convergent.

PROOF: As \hat{f} is integrable, we have

$$\begin{aligned} \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi)e^{-i\xi x} d\xi - \frac{1}{2\pi} \int_{-R}^S \hat{f}(\xi)e^{-i\xi x} d\xi \right| &= \frac{1}{2\pi} \left| \int_{\xi < -R \text{ or } \xi > S} \hat{f}(\xi)e^{-i\xi x} d\xi \right| \\ &\leq \frac{1}{2\pi} \int_{\xi < -R \text{ or } \xi > S} |\hat{f}(\xi)| d\xi \rightarrow 0 \quad \text{as } R, S \rightarrow \infty, \end{aligned}$$

and so the Fourier inversion integral converges uniformly. We now need to show that it actually converges to f .

By the spotlight theorem, the right-hand side of (3.2) converges to $f(x)$ as $\sigma \rightarrow 0$. On the other hand,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\xi)\phi(\sigma\xi)e^{-i\xi x} d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi)e^{-\sigma^2 \xi^2/2} e^{-i\xi x} d\xi$$

so that by comparison with the Fourier inversion integral we obtain

$$\begin{aligned} & \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{-\sigma^2 \xi^2/2} e^{-i\xi x} d\xi - \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{-i\xi x} d\xi \right| \\ &= \frac{1}{2\pi} \left| \int_{-\infty}^{\infty} \widehat{f}(\xi) (e^{-\sigma^2 \xi^2/2} - 1) e^{-i\xi x} d\xi \right| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{f}(\xi)| \cdot |e^{-\sigma^2 \xi^2/2} - 1| d\xi. \quad (3.3) \end{aligned}$$

To show that the integral on the right is small we resort to a, by now standard, partitioning of the range of integration. As \widehat{f} is integrable, the improper integral $\int_{-\infty}^{\infty} |\widehat{f}(\xi)| d\xi$ converges, say, to a finite value M . For any fixed $\epsilon > 0$ we may hence select $\Omega = \Omega(\epsilon)$ sufficiently large so that $\int_{|\xi|>\Omega} |\widehat{f}(\xi)| d\xi < \epsilon$. In view of the monotonicity of the exponential function, we now select σ sufficiently small so that $|e^{-\sigma^2 \xi^2/2} - 1| < \epsilon$ whenever $|\xi| \leq \Omega$; for $|\xi| > \Omega$ we can afford to be more cavalier with the bound $|e^{-\sigma^2 \xi^2/2} - 1| \leq 2$. We may then upper bound the right-hand side of (3.3) by

$$\frac{\epsilon}{2\pi} \int_{|\xi| \leq \Omega} |\widehat{f}(\xi)| d\xi + \frac{2}{2\pi} \int_{|\xi| > \Omega} |\widehat{f}(\xi)| d\xi \leq \frac{(M+2)\epsilon}{2\pi}$$

and as ϵ may be chosen arbitrarily small the bound can be made as small as desired. It follows that, for each x ,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{-\sigma^2 \xi^2/2} e^{-i\xi x} d\xi \rightarrow \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{-i\xi x} d\xi$$

as $\sigma \rightarrow 0$. As both sides of (3.2) go to distinct limits as $\sigma \rightarrow 0$ the limits must be equal and the theorem is proved. ▶

A key consequence of the inversion formula is the famous identity linked with Parseval. We shall content ourselves with a version for continuous functions. For a discrete formulation in L^2 theory, see Section XXI.3.

PARSEVAL'S EQUATION *Suppose f , g , \widehat{f} , and \widehat{g} are continuous and integrable. Then*

$$\int_{-\infty}^{\infty} f(x) g(x)^* dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi) \widehat{g}(\xi)^* d\xi \text{ and } \int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2 d\xi. \quad (3.4)$$

PROOF: In view of the Fourier formula and its inverse, all of the functions f , g , \widehat{f} , and \widehat{g} are bounded (and indeed uniformly continuous). It follows in particular that the function $f(x)g(x)^*$ is continuous, bounded, and integrable and it is legitimate to change the order of integration to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) g(x)^* dx &= \int_{-\infty}^{\infty} f(x) \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{g}(\xi) e^{-i\xi x} d\xi \right)^* dx \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{g}(\xi)^* \int_{-\infty}^{\infty} f(x) e^{i\xi x} dx d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{g}(\xi)^* \widehat{f}(\xi) d\xi. \end{aligned}$$

The special case $f = g$ simplifies the expression to the second of the claimed identities. ▶

In the second of the two forms in (3.4) the identity is sometimes called *Plancherel's equation*. If we identify f with a physical waveform then the equation may be interpreted as a statement of energy conservation.

4 An idea of Markov

With notation as in Chapter V, let S_n denote the number of successes in n tosses of a fair coin and let $R_n(t) = r_1(t) + \dots + r_n(t)$ be the sum of the first n Rademacher functions. Chebyshev's argument leading up to (V.6.2) shows that

$$P\{|S_n - n/2| \geq \Delta\} = \lambda\{t : |R_n(t)| \geq 2\Delta\} \leq \frac{n}{4\Delta^2}.$$

The upper bound becomes vanishingly small if $\Delta \gg \sqrt{n}$ which suggests that when Δ is of the order of \sqrt{n} matters become interesting: Figure 4 highlights probabilities in the critical central region when $n = 100$. The algebra is sim-

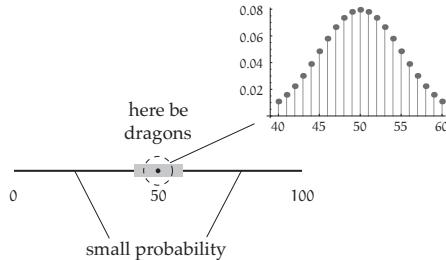


Figure 4: Where the wild things are.

plest if we consider the deviation from $n/2$ in units of $\sqrt{n}/2$, that is to say, we consider the (dimensionless) deviation $\Delta/\frac{\sqrt{n}}{2}$, and we accordingly define the normalised variable

$$S_n^* = \frac{S_n - n/2}{\sqrt{n}/2} = \frac{2S_n - n}{\sqrt{n}}$$

which allows us to look at matters in the proper scale. For any $a < b$, by the correspondence between coin tosses and binary digits, the event that $a < S_n^* < b$ may be placed in one-to-one correspondence with the set of points t in the unit interval for which $a < [2(z_1(t) + \dots + z_n(t)) - n]/\sqrt{n} < b$ or, equivalently, $-b < \frac{1}{\sqrt{n}} \sum_{k=1}^n (1 - 2z_k(t)) < -a$. Accordingly,

$$P\{a < S_n^* < b\} = \lambda\left\{t : -b < \frac{R_n(t)}{\sqrt{n}} < -a\right\} = \int_0^1 1_{\mathbb{A}}(t) dt,$$

where \mathbb{A} now represents the set of points t on which $-b < R_n(t)/\sqrt{n} < -a$. The expression on the right becomes slightly more expressive if we introduce the nonce function

$$f(x) = \text{rect}\left(\frac{x + (a + b)/2}{b - a}\right) = \begin{cases} 1 & \text{if } -b < x < -a, \\ 0 & \text{otherwise,} \end{cases} \quad (4.1)$$

as we may now write

$$\mathbf{P}\{a < S_n^* < b\} = \int_0^1 f\left(\frac{R_n(t)}{\sqrt{n}}\right) dt$$

and it becomes clear that the evaluation of the right-hand side depends critically on the interaction of the rectangular function f with the Rademacher functions.

The Fourier transform of f is readily computed by direct integration to be

$$\hat{f}(\xi) = \frac{e^{-i\xi a} - e^{-i\xi b}}{i\xi},$$

the right-hand side to be interpreted as equal to $b - a$ when $\xi = 0$. Fourier inversion now says that we may recover f , at least at points of continuity, via

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{-i\xi x} d\xi \quad (x \notin \{-b, -a\}).$$

The reader may object that $\hat{f}(\xi)$ has heavy tails decaying like the reciprocal of ξ , hence is not (absolutely) integrable; for the Fourier inversion integral to converge then it appears that it will require a fairly delicate cancellation of terms at infinity. True. But we won't pause to pull this particular chestnut out of the fire as we will soon be able to finesse the need to deal with the baulky \hat{f} altogether.

Pressing forward accordingly to see where this line of thought leads, the Fourier inversion formula for f suggests the intriguing sequence of steps

$$\begin{aligned} \mathbf{P}\{a < S_n^* < b\} &= \frac{1}{2\pi} \int_0^1 \int_{-\infty}^{\infty} \hat{f}(\xi) \exp\left(\frac{-i\xi R_n(t)}{\sqrt{n}}\right) d\xi dt \\ &\stackrel{(\dagger)}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \int_0^1 \exp\left(\frac{-i\xi R_n(t)}{\sqrt{n}}\right) dt d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \int_0^1 \prod_{k=1}^n \exp\left(\frac{-i\xi r_k(t)}{\sqrt{n}}\right) dt d\xi \\ &\stackrel{(\ddagger)}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \prod_{k=1}^n \int_0^1 \exp\left(\frac{-i\xi r_k(t)}{\sqrt{n}}\right) dt d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \cos\left(\frac{\xi}{\sqrt{n}}\right)^n d\xi, \end{aligned} \quad (4.2)$$

the interchange of integrals in the step labelled (\dagger) occasioning no difficulty, assuming convergence of the Fourier inversion integral, as $R_n(t)$ only assumes a finite number of values, and the step labelled (\ddagger) justified by the independence of the binary digits. We are hence led to consider the behaviour of

$$\cos\left(\frac{\xi}{\sqrt{n}}\right)^n = \exp\left[n \log\left\{\cos\left(\frac{\xi}{\sqrt{n}}\right)\right\}\right].$$

By Taylor's theorem, for any z , we have $\cos(z) = 1 - \frac{1}{2}z^2 + \frac{1}{24}\cos(\theta)z^4$ where $\theta = \theta(z)$ is a point between 0 and z . As the cosine is bounded in absolute value by 1, it follows that

$$\cos\left(\frac{\xi}{\sqrt{n}}\right) = 1 - \frac{\xi^2}{2n} + \zeta_1 \frac{\xi^4}{n^2}$$

where $\zeta_1 = \zeta_1(n, \xi)$ is bounded in absolute value by $1/24$. The final two terms on the right are asymptotically subdominant. Indeed, if we write

$$x = \frac{\xi^2}{2n} - \zeta_1 \frac{\xi^4}{n^2} = \frac{\xi^2}{2n} \left(1 - 2\zeta_1 \frac{\xi^2}{n}\right),$$

then it becomes clear that $0 < x < \xi^2/n$ for all sufficiently large n with the upper bound becoming less than 1, eventually; the algebra is simplified if we select $n > \xi^2\sqrt{2}/(\sqrt{2}-1)$ in which case $0 < x < \xi^2/n < 1 - 1/\sqrt{2}$. Another application of Taylor's theorem, for the natural logarithm this time, shows now that, for some $y = y(x)$ between 0 and x , we have $g(x) = \log(1-x) = -x + g''(y)x^2/2$, where $g''(y) = -(1-y)^{-2}$ is bounded in absolute value by 2 when $n > \xi^2\sqrt{2}/(\sqrt{2}-1)$. It follows that $\log(1-x) = -x + \zeta_2 x^2$ for some $\zeta_2 = \zeta_2(x) = \zeta_2(n, \xi)$ bounded in absolute value by 1, eventually, as n becomes sufficiently large. Putting the pieces together, for all sufficiently large n , we have

$$\cos\left(\frac{\xi}{\sqrt{n}}\right)^n = \exp(n \log(1-x)) = \exp(n(-x + \zeta_2 x^2)) = \exp\left(-\frac{\xi^2}{2} + \zeta_3 \frac{\xi^4}{n}\right)$$

where $\zeta_3 = \zeta_3(n, \xi)$ is bounded in absolute value by $25/24$. Allowing n to tend to infinity on both sides, we see that $\cos(\xi/\sqrt{n})^n$ converges pointwise to the Gaussian characteristic function $\hat{\phi}(\xi) = e^{-\xi^2/2}$. In the excitement of the moment we should not overlook the fact that the convergence is *not* uniform—as ξ increases in absolute value, ever larger values of n are needed to get a good approximation.

If, as $n \rightarrow \infty$, either side of (4.2) converges to a limit, then both do and the limits will be the same. By taking formal limits of both sides, the great Russian mathematician Andrei Markov was hence led to entertain the following speculative sequence of steps:

$$\begin{aligned} \lim_n \mathbf{P}\{a < S_n^* < b\} &= \lim_n \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \cos\left(\frac{\xi}{\sqrt{n}}\right)^n d\xi \\ &\stackrel{?}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \lim_n \cos\left(\frac{\xi}{\sqrt{n}}\right)^n d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \hat{\phi}(\xi) d\xi \stackrel{(i)}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) \hat{\phi}(\xi)^* d\xi \\ &\stackrel{(ii)}{=} \int_{-\infty}^{\infty} f(x) \phi(x)^* dx \stackrel{(iii)}{=} \int_{-\infty}^{\infty} f(x) \phi(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-b}^{-a} e^{-x^2/2} dx \stackrel{(iv)}{=} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx. \end{aligned} \tag{4.3}$$

Step (iv) is justified because ϕ is an even function, steps (i) and (iii) follow because $\hat{\phi}$ and ϕ are real-valued functions, and step (ii) follows by Parseval's equation. The only step that needs justification is the one marked ? in which the limit and the integral are interchanged. To be absolutely secure we will need a result along the lines of the one sketched below; the reader can skip the proof if she is not feeling pernickety.

LEMMA Suppose $\{\gamma_n\}$ is a sequence of continuous and uniformly bounded functions converging pointwise to a continuous limit function γ , the convergence uniform over every closed and bounded interval. Suppose ρ is an integrable function. Then

$$\lim_n \int_{-\infty}^{\infty} \rho(\xi) \gamma_n(\xi) d\xi = \int_{-\infty}^{\infty} \rho(\xi) \lim_n \gamma_n(\xi) d\xi = \int_{-\infty}^{\infty} \rho(\xi) \gamma(\xi) d\xi.$$

PROOF: We begin by fixing any $\epsilon > 0$. As ρ is integrable we may then select $\Omega = \Omega(\epsilon)$ such that $\int_{|\xi| > \Omega} |\rho(\xi)| d\xi < \epsilon$. Accordingly, we may divide and conquer to obtain

$$\begin{aligned} \left| \int_{-\infty}^{\infty} \rho(\xi) \gamma_n(\xi) d\xi - \int_{-\infty}^{\infty} \rho(\xi) \gamma(\xi) d\xi \right| &= \left| \int_{-\infty}^{\infty} \rho(\xi) [\gamma_n(\xi) - \gamma(\xi)] d\xi \right| \\ &\leq \int_{|\xi| \leq \Omega} |\rho(\xi)| \cdot |\gamma_n(\xi) - \gamma(\xi)| d\xi + \int_{|\xi| > \Omega} |\rho(\xi)| \cdot |\gamma_n(\xi) - \gamma(\xi)| d\xi. \end{aligned} \quad (4.4)$$

As $\{\gamma_n\}$ converges uniformly to γ in the interval $[-\Omega, \Omega]$, we will have $|\gamma_n(\xi) - \gamma(\xi)| < \epsilon$ for all $\xi \in [-\Omega, \Omega]$ eventually, say for $n \geq N(\epsilon)$. Writing $M_1 = \int_{-\infty}^{\infty} |\rho(\xi)| d\xi$, the first term on the right in (4.4) is hence seen to be bounded above by $M_1 \epsilon$, eventually, for all sufficiently large n . Furthermore, as the uniformly bounded sequence $\{\gamma_n\}$ converges pointwise to γ , it follows that there exists some M_2 such that $|\gamma_n|, |\gamma| \leq M_2$ for all n . The second term on the right in (4.4) is hence bounded above by $M_2 \epsilon$. It follows that

$$\left| \int_{-\infty}^{\infty} \rho(\xi) \gamma_n(\xi) d\xi - \int_{-\infty}^{\infty} \rho(\xi) \gamma(\xi) d\xi \right| \leq (M_1 + M_2) \epsilon$$

eventually. As the positive ϵ may be chosen as small as desired, it follows that the limit and integral may be interchanged as asserted. ▶

The reader may wonder if the conditions are too onerous and whether, perhaps, the condition that ρ be absolutely integrable can be relaxed. The following example may serve to convince her otherwise.

A CONTRARIAN EXAMPLE ON EXCHANGING LIMITS AND INTEGRALS Beginning with the unit triangle function $\Delta(\xi) = \max\{0, 1 - |\xi|\}$, let $\gamma_n(\xi) = \Delta(\xi - n)$ denote its translated version with the origin placed at the point n . It is clear that, for each n , γ_n is continuous and bounded (by 1) and $\int_{-\infty}^{\infty} \gamma_n(\xi) d\xi = \int_{-\infty}^{\infty} \Delta(\xi) d\xi = 1$. As γ_n has support in the interval $(n - 1, n + 1)$, for any fixed ξ_0 it follows that $\gamma_n(\xi_0) = 0$ whenever $n \geq \xi_0 + 1$ and, *a fortiori*, $\gamma_n(\xi_0) \rightarrow 0$ as $n \rightarrow \infty$. Consequently, $\{\gamma_n(\xi)\}$ converges pointwise to the identically zero function $\gamma(\xi) = 0$, the convergence uniform over every interval $(-\infty, \xi_0]$, *a fortiori* over every closed and bounded interval. Identifying ρ with

the function that takes the constant value 1 everywhere, we hence have the upsetting conclusion that, even in this transparent setting,

$$1 = \lim_n \int_{-\infty}^{\infty} \rho(\xi) \gamma_n(\xi) d\xi \neq \int_{-\infty}^{\infty} \rho(\xi) \lim_n \gamma_n(\xi) d\xi = 0.$$

In spite of the simplicity of the functions γ_n and ρ , the difficulty in interchanging limit and integration arises here because the function $\rho(\xi) = 1$ is not integrable. ▶

So, where does this leave us? The functions $\gamma_n(\xi) = \cos(\xi/\sqrt{n})^n$ are continuous and uniformly bounded, the sequence converging pointwise and uniformly over every closed and bounded interval to the continuous function $\hat{\phi}$. So far so good. But when we consider the function $\widehat{f}(\xi)$ we hit a stumbling block: as the function decays as the reciprocal of its argument we are faced with the inconvenient fact that \widehat{f} is not integrable.

Markov could not convince himself that the limit and the integral could be interchanged in view of examples like the contrarian one above and eventually abandoned the approach. Some twenty years later the probabilist Paul Lévy resuscitated the idea by an ingenious approximation argument.

5 Lévy suggests a thin sandwich, de Moivre redux

The heavy tails of the function $\widehat{\text{rect}}(\xi) = \text{sinc}(\xi/2)$ leading to its non-integrability may be directly traced to the sharp corners of the rectangular function $\text{rect}(x)$ at the points of discontinuity $x = \pm 1/2$. Lévy realised that we can improve the tail behaviour of the Fourier transform by making the function a bit smoother. The simplest approach is to smooth the rectangular corners to form a trapezoidal graph as shown in Figure 3. As before, for each non-zero τ , we define

$$\text{trap}_\tau(x) = \frac{1/2 + \tau}{\tau} \Delta\left(\frac{x}{1/2 + \tau}\right) - \frac{1/2}{\tau} \Delta\left(\frac{x}{1/2}\right),$$

where $\Delta(x) = \max[0, 1 - |x|]$ represents the unit triangular function. From Example 2.3 we may conclude that $|\widehat{\text{trap}}_\tau(\xi)| \leq 2/(\tau \xi^2)$ if $\xi \neq 0$ so that, as Lévy anticipated, $\widehat{\text{trap}}_\tau(\xi)$ is integrable for every $\tau \neq 0$.

For any $\epsilon > 0$, it is clear by inspection of the graphs of the functions that $\text{trap}_{-\epsilon}(x) \leq \text{rect}(x) \leq \text{trap}_\epsilon(x)$. Starting with the function f defined in (4.1), if we now define the shifted and axis-scaled smoothed variants

$$f_{\pm}(x) = \text{trap}_{\pm\epsilon}\left(\frac{x + (a + b)/2}{b - a}\right),$$

then $f_-(x) \leq f(x) \leq f_+(x)$ and the graphs of the trapezoidal functions f_- and f_+ neatly sandwich that of the rectangular function f . It follows that

$$\int_0^1 f_-\left(\frac{R_n(t)}{\sqrt{n}}\right) dt \leq \int_0^1 f\left(\frac{R_n(t)}{\sqrt{n}}\right) dt \leq \int_0^1 f_+\left(\frac{R_n(t)}{\sqrt{n}}\right) dt.$$

As the trapezoidal functions f_{\pm} are so much better behaved than the rectangular function f , the integrals on either side are simple to evaluate: indeed, with f replaced by f_{\pm} , the string of steps in (4.3) is impeccable as \widehat{f}_{\pm} is now integrable and we obtain

$$\begin{aligned} \lim_n \int_0^1 f_{\pm} \left(\frac{R_n(t)}{\sqrt{n}} \right) dt &= \lim_n \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}_{\pm}(\xi) \cos \left(\frac{\xi}{\sqrt{n}} \right)^n d\xi \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}_{\pm}(\xi) \lim_n \cos \left(\frac{\xi}{\sqrt{n}} \right)^n d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}_{\pm}(\xi) \widehat{\phi}(\xi) d\xi = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}_{\pm}(\xi) \widehat{\phi}(\xi)^* d\xi \\ &= \int_{-\infty}^{\infty} f_{\pm}(x) \phi(x)^* dx = \int_{-\infty}^{\infty} f_{\pm}(x) \phi(x) dx. \end{aligned}$$

As f_{\pm} differs from f (and by no more than one unit) only in intervals of total length $2(b-a)\epsilon$, by selecting ϵ small enough, we may make the integral on the right differ from $\int_{-\infty}^{\infty} f(x)\phi(x) dx$ in as little as we wish. To make the argument completely formal needs little more than a whisk with a formalism brush. As $\phi(x) \leq 1$ for all x (crude bounds suffice here),

$$\left| \int_{-\infty}^{\infty} f_{\pm}(x) \phi(x) dx - \int_{-\infty}^{\infty} f(x) \phi(x) dx \right| = \left| \int_{-\infty}^{\infty} (f_{\pm}(x) - f(x)) \phi(x) dx \right| \leq 2(b-a)\epsilon.$$

We have managed to sandwich the desired probability,

$$\begin{aligned} -2(b-a)\epsilon + \int_{-\infty}^{\infty} f(x) \phi(x) dx &\leq \int_{-\infty}^{\infty} f_{-}(x) \phi(x) dx \\ &= \lim_n \int_0^1 f_{-} \left(\frac{R_n(t)}{\sqrt{n}} \right) dt \leq \lim_n \int_0^1 f \left(\frac{R_n(t)}{\sqrt{n}} \right) dt \leq \lim_n \int_0^1 f_{+} \left(\frac{R_n(t)}{\sqrt{n}} \right) dt \\ &= \int_{-\infty}^{\infty} f_{+}(x) \phi(x) dx \leq \int_{-\infty}^{\infty} f(x) \phi(x) dx + 2(b-a)\epsilon, \end{aligned}$$

or, more compactly,

$$\left| \lim_n \int_0^1 f \left(\frac{R_n(t)}{\sqrt{n}} \right) dt - \int_{-\infty}^{\infty} f(x) \phi(x) dx \right| \leq 2(b-a)\epsilon.$$

As ϵ may be chosen as small as desired, it must follow that

$$\lim_n \int_0^1 f \left(\frac{R_n(t)}{\sqrt{n}} \right) dt = \int_{-\infty}^{\infty} f(x) \phi(x) dx = \int_{-b}^{-a} \phi(x) dx = \int_a^b \phi(x) dx,$$

the final step a consequence of the fact that ϕ is an even function. Our findings are so significant that we should promptly capture them in a theorem and a slogan.

DE MOIVRE'S THEOREM *Let S_n be the number of successes in n tosses of a fair coin and let $S_n^* = (S_n - \frac{1}{2}n) / (\frac{1}{2}\sqrt{n})$ represent the centred and properly scaled number of successes. Then*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{a < S_n^* < b\} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx \quad (5.1)$$

for every choice of a and b with $a < b$.

In the language of probability, we say that S_n^* converges in distribution to the normal. This notion of convergence is very different from the ideas of convergence in probability or convergence a.e. we have encountered hitherto which, at least, viewed S_n in its more familiar rôle as a sequence *qua* sequence of numbers. When we talk about convergence in distribution, however, we are making a statement about the probability law that the sequence follows asymptotically, not its individual limiting value.

If we set $X_k = 1$ if the k th toss is a success and $X_k = 0$ if the k th toss is a failure we may identify S_n as the sum of the variables X_1 through X_n and S_n^* as a suitably normalised sum by centring and scaling. In this light we may represent our result via the following

SLOGAN *The sum of independent perturbations obeys a normal law.*

A simple numerical example may serve to illustrate the theorem's sway.

EXAMPLE: *Random walks.* A particle starts at the origin and at each succeeding time instant moves randomly one unit left or right with equal probability $1/2$. The particle is then said to execute a random walk on the line. If T_n denotes the particle's position after n time epochs, then $S_n = (T_n + n)/2$ represents the number of successes in n tosses of a coin. It follows that

$$\mathbf{P}\{|T_n| \geq 3\sqrt{n}\} = \mathbf{P}\left\{\left|\frac{S_n - n/2}{\sqrt{n}/2}\right| \geq 3\right\} \approx 2 \int_3^\infty \phi(x) dx = 0.0027\dots$$

For a walk over 10,000 steps, for instance, 99.7% of all random walks return to within 300 steps of the origin; the huge space between 300 and 10,000 is filled by only 0.3% of the walks. Informally speaking, most random walks return to a vicinity of the point of origin. ►

The beauty and importance of theorems like those of de Moivre is at once both conceptual and computational. At a fundamental level, by articulating a simple limit law the theorem vastly improves our understanding of a phenomenon which, on closer examination, appears to be very complex. And the existence of such a simple limit law opens up the possibility that results of this stripe may be obtainable for other phenomena; and, indeed, this was to prove to be the case—Markov's idea was to open up the floodgates for general limit laws for a wide range of random processes. If the reader is not convinced by such

specious philosophical arguments she may perhaps be swayed by an argument at a practical level: the theorem provides a principled justification for a computationally simple approximation to a complex procedure. Indeed, the probability in question deals with the evaluation of the sum $P\{a < S_n^* < b\} = \sum \binom{n}{k} 2^{-n}$ where k varies over the range $\frac{1}{2}(n - a\sqrt{n}) < k < \frac{1}{2}(n + b\sqrt{n})$ and the reader will only need to try to compute the sum by direct calculation for n equal to, say, 10^{12} to appreciate the computational simplicity of the approximation (5.1). Of course, the utility of the approximation for any given n will depend on how much of an error is made and, to be absolutely safe, we should provide error bounds. While this is possible, we won't pause to do so here relying instead upon Körner's bromide that a principled answer (for which we have a theoretical basis for belief) in a small number of steps is in any case better than no answer at all after a large number of steps.

Abraham de Moivre discovered his theorem in 1733 by direct combinatorial methods. While elementary, these arguments tend to obscure the ubiquitous nature of the result and its genuine power—features that Markov's method lays out in sharp relief. De Moivre's theorem was extended by Pierre Simon Marquis de Laplace in 1812 to tosses of bent coins—Problem 11 sketches how one may adapt our proof to this setting—and thence generalised by degrees culminating in a famous theorem of J. W. Lindeberg in 1922. As a wide range of phenomena may be modelled as the sum of independent perturbations, the theorem explains the remarkable ubiquity of the normal curve in theory and practice. Indeed, it is hard to overstate its continuing impact in diverse applications from all areas of science and engineering. Armed with more background, we will return to the subject and Lindeberg's proof in Chapter XX.



6 A local limit theorem

De Moivre's theorem tells us how the cumulative distribution of successes in repeated tosses of a coin varies asymptotically with the number of tosses. To examine the fine structure of the distribution of successes we turn to the representations (V.5.2, V.5.2').

In a succession of tosses of a fair coin we expect that the most likely outcome is that half the coins turn heads. Indeed, a consideration of the ratio of successive terms in (V.5.2') shows that

$$\frac{b_n(k)}{b_n(k-1)} = \frac{\binom{n}{k} 2^{-n}}{\binom{n}{k-1} 2^{-n}} = \frac{n-k+1}{k} \quad (6.1)$$

whence $b_n(k) \geq b_n(k-1)$ if, and only if, $k \leq (n+1)/2$. Writing

$$m = \lceil n/2 \rceil = \begin{cases} n/2 & \text{if } n \text{ is even,} \\ (n+1)/2 & \text{if } n \text{ is odd,} \end{cases}$$

it follows that, for each n , the probabilities $b_n(k)$ increase monotonically with k up till $k = m$ and decrease monotonically thereafter; the most likely eventuality corresponds

hence to exactly m successes and it will be convenient to centre the problem and consider the nature of the sequence $b_n(m+k)$ for $k = 0, \pm 1, \pm 2, \dots$. We introduce the temporary notation $\beta_k = b_n(m+k)$ to focus more clearly on the dependence of the sequence on the deviations k from the centre.

We begin by estimating the central term $\beta_0 = b_n(m)$ when n is large. While a direct combinatorial attack through Stirling's formula is possible starting with the expression (V.5.2'), it is more in keeping with the tenor of this chapter to start instead with the representation (V.5.2). As a starting point for our investigation we hence begin with

$$\beta_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ix(2m-n)} \cos(x)^n dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos((2m-n)x) \cos(x)^n dx,$$

the second step following by writing $e^{-ix(2m-n)} = \cos((2m-n)x) - i \sin((2m-n)x)$ and observing that the integrand $\sin((2m-n)x) \cos(x)^n$ for the contribution to the imaginary part of β_0 is an odd function of x and hence integrates to zero. As $2m-n=0$ for n even and $2m-n=1$ for n odd, in all cases the expression devolves to the integral of cosine to an *even* power over an interval of length 2π . We may further compact the range of integration by the observation that the cosine to an even power has period π . As the integral of a periodic function over any interval of length equal to its period is invariant with respect to the location of the interval, the integral at hand from $-\pi$ to π yields exactly twice the contribution from any one period which, for definiteness, we may take from $-\pi/2$ to $\pi/2$. It follows that

$$\beta_0 = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \cos(x)^\nu dx \quad (6.2)$$

where $\nu = n$ if n is even and $\nu = n+1$ if n is odd. The expressions for even and odd cases have the same form and for definiteness we consider hence that n is even.

In broad brush strokes, the key to the estimation of the integral expression for β_0 is the observation that $\cos(x)^n$ is highly concentrated at the origin so that most of the contribution to the integral comes from a vicinity of the origin. To put the observation to good use, select any $0 < \epsilon < \pi/2$. As $\cos(x)$ is positive everywhere in the interval $-\pi/2 < x < \pi/2$ and, moreover, decays monotonically away from the origin in this interval, the contribution to the integral from the region $\epsilon < |x| < \pi/2$ may be bounded by

$$0 < \frac{1}{\pi} \int_{\epsilon < |x| < \pi/2} \cos(x)^n dx < \cos(\epsilon)^n = e^{-Cn}. \quad (6.3)$$

Here $C = C(\epsilon) = -\log \cos \epsilon > 0$. The contribution to the integral away from the origin will hence be exponentially subdominant and we are left with evaluating the central contribution

$$\frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \cos(x)^n dx = \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp(n \log \cos(x)) dx.$$

It is easy to see that the function $f(x) = \log \cos x$ is ≤ 0 everywhere and achieves its maximum value in the interval $[-\epsilon, \epsilon]$ at the point $x = 0$. Indeed, repeated differentiation shows that $f'(x) = -\sin(x)/\cos(x)$, $f''(x) = -1/\cos(x)^2$, $f'''(x) = -2 \sin(x)/\cos(x)^3$, and $f''''(x) = -2(1 + 2 \sin(x)^2)/\cos(x)^4$. It follows that $f(x)$, $f''(x)$, and $f'''(x)$ are all ≤ 0 and bounded in the interval $[-\epsilon, \epsilon]$. Moreover, by setting f' to zero we verify that,

in this interval, f achieves its unique maximum value of 0 at $x = 0$ and $f(0) = f'(0) = f'''(0) = 0$ while $f''(0) = -1$. By Taylor's theorem truncated to five terms, we have

$$f(x) = f(0) + f'(0)x + f''(0)\frac{x^2}{2} + f'''(0)\frac{x^3}{3!} + f''''(\theta)\frac{x^4}{4!}$$

for some $\theta = \theta(x)$ between 0 and x . It follows that

$$f(x) = -\frac{x^2}{2} - \frac{1+2\sin(\theta)^2}{12\cos(\theta)^4}x^4 = -\frac{x^2}{2}(1+\zeta x^2)$$

for some $\zeta = \zeta(x)$ satisfying $0 < \zeta < 1$ for all x in a sufficiently small interval $|x| < \epsilon$. We may, for instance, take $\epsilon < \arccos(2^{-1/4}) = 0.57 \dots$. Simple bounds suffice here: as $0 \leq \zeta x^2 \leq \epsilon^2 \leq \epsilon$, we have $-(1+\epsilon)x^2/2 \leq f(x) \leq -(1-\epsilon)x^2/2$ whenever $|x| \leq \epsilon$. It follows that

$$\frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp\left(-\frac{1}{2}(1+\epsilon)nx^2\right) dx \leq \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp(nf(x)) dx \leq \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp\left(-\frac{1}{2}(1-\epsilon)nx^2\right) dx. \quad (6.4)$$

The integrals on either side look like Gaussian integrals. Indeed, by the change of variable $u = x\sqrt{n(1+\epsilon)}$ for the lower bound we obtain

$$\begin{aligned} \frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp\left(-\frac{1}{2}(1+\epsilon)nx^2\right) dx &= \frac{1}{\pi(1+\epsilon)^{1/2}n^{1/2}} \int_{-\sqrt{2cn}}^{\sqrt{2cn}} e^{-u^2/2} du \\ &= \sqrt{\frac{2}{\pi n}} (1+\epsilon)^{-1/2} \int_{-\sqrt{2cn}}^{\sqrt{2cn}} \phi(u) du \end{aligned}$$

where $c = c(\epsilon) = \epsilon^2(1+\epsilon)/2$. As ϕ is even, by Lemmas 1.1 and 1.3 we obtain

$$\int_{-\sqrt{2cn}}^{\sqrt{2cn}} \phi(u) du = 1 - 2 \int_{\sqrt{2cn}}^{\infty} \phi(u) du \geq 1 - e^{-cn}$$

and it follows that

$$\frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp\left(-\frac{1}{2}(1+\epsilon)nx^2\right) dx \geq \sqrt{\frac{2}{\pi n}} (1+\epsilon)^{-1/2} (1 - e^{-cn}). \quad (6.5)$$

Similar but easier bounds suffice for the upper bound in (6.4). We have

$$\frac{1}{\pi} \int_{-\epsilon}^{\epsilon} \exp\left(-\frac{1}{2}(1-\epsilon)nx^2\right) dx \leq \frac{1}{\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}(1-\epsilon)nx^2\right) dx = \sqrt{\frac{2}{\pi n}} (1-\epsilon)^{-1/2}. \quad (6.6)$$

Putting together the estimates from (6.3,6.4,6.5,6.6), we have

$$\sqrt{\frac{2}{\pi n}} (1+\epsilon)^{-1/2} (1 - e^{-cn}) \leq \underbrace{\frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \cos(x)^n dx}_{\beta_0} \leq \sqrt{\frac{2}{\pi n}} (1-\epsilon)^{-1/2} + e^{-cn}.$$

As, for all sufficiently small ϵ , and certainly for all ϵ in the range $(0, 0.618)$, we have $1 - \epsilon \leq (1 + \epsilon)^{-1/2} \leq 1$ and $(1 - \epsilon)^{-1/2} \leq 1 + \epsilon$, by dividing throughout by $\sqrt{2/\pi n}$, we obtain

$$1 - \epsilon - e^{-cn} \leq \frac{\beta_0}{\sqrt{2/\pi n}} \leq 1 + \epsilon + \sqrt{\frac{\pi}{2}} \sqrt{n} e^{-cn}$$

where we have cheerfully worsened the bounds to keep the small irritant terms in simplest terms. For any choice of $\epsilon > 0$, we may select n sufficiently large so that both exponential terms in the bounds above are no larger than ϵ . Consequently,

$$1 - 2\epsilon \leq \frac{\beta_0}{\sqrt{2/\pi n}} \leq 1 + 2\epsilon,$$

eventually, for large enough values of n . As $\epsilon > 0$ may be chosen arbitrarily small, the upper and lower bounds both differ from 1 in as small an amount as one pleases. The case of odd n is handled by simply replacing n by $n + 1$ in the expression above. But as

$$\frac{\sqrt{2/\pi n}}{\sqrt{2/\pi(n+1)}} = \left(1 + \frac{1}{n}\right)^{1/2}$$

is as close to one as desired for all sufficiently large n , we may replace $\sqrt{2/\pi(n+1)}$ by $\sqrt{2/\pi n}$ incurring a relative error that is as small as desired eventually. It follows that we need not make a distinction between even and odd cases asymptotically and we obtain the desired asymptotic estimate

$$\beta_0 \sim \sqrt{\frac{2}{\pi n}} \quad (n \rightarrow \infty) \tag{6.7}$$

where the asymptotic equivalence relation \sim is to be taken to mean that the ratio of the two sides tends to one as $n \rightarrow \infty$.

The approach outlined here for the asymptotic evaluation of (6.2) is of broad general utility. It may be pressed into service whenever one wishes to evaluate an integral of the form $\int_{\mathbb{I}} e^{nf(x)} dx$ where \mathbb{I} is a finite or infinite interval. We expect in such cases that most of the contribution to the integral arises from a small neighbourhood of the point where f achieves its maximum. This is the essence of *Laplace's method of integration*. Section XIV.7 provides another case in point for the reader who would like to reinforce her understanding of the method in another setting of great analytical importance.

We now turn to estimating β_k when n is large. It is clear from either of the representations (V.5.2, V.5.2') that $\beta_{-k} = \beta_k$ and we may as well consider $k \geq 0$. Rewriting (6.1), it is clear that

$$\beta_k = \frac{n-m-k+1}{m+k} \beta_{k-1}$$

so that a quick process of induction yields

$$\beta_k = \frac{n-m-k+1}{m+k} \cdot \frac{n-m-k+2}{m+k-1} \cdots \frac{n-m-1}{m+2} \cdot \frac{n-m}{m+1} \cdot \beta_0.$$

The expression on the right takes on slightly different forms depending on whether n is even or odd. We may combine the cases by introducing the nuisance variable $\alpha = \alpha(n)$

which takes value 0 if n is even and value 1 if n is odd. The generic term in the product on the right is then of the form

$$\frac{n-m-j}{m+j+1} = \frac{\frac{n}{2} - \frac{\alpha}{2} - j}{\frac{n}{2} + \frac{\alpha}{2} + j + 1} = \frac{\frac{n+1}{2} - (j + \frac{\alpha}{2} + \frac{1}{2})}{\frac{n+1}{2} + (j + \frac{\alpha}{2} + \frac{1}{2})}$$

and by dividing both numerator and denominator in each such fraction by $(n+1)/2$ it follows that

$$\beta_k = \beta_0 \prod_{j=0}^{k-1} \frac{1-x_j}{1+x_j} = \beta_0 \exp \left(\sum_{j=0}^{k-1} \log \frac{1-x_j}{1+x_j} \right), \quad (6.8)$$

where we write $x_j = (2j+\alpha+1)/(n+1)$ to keep the burgeoning notation under control. From the Taylor expansion for the logarithm it is clear that

$$\log \frac{1-x}{1+x} = -2x - \frac{2}{3}x^3 - \frac{2}{5}x^5 - \dots = -2x - \frac{2}{3}x^3(1 + \frac{3}{5}x^2 + \frac{3}{7}x^4 + \dots) = -2x - \zeta(x)x^3$$

where, by comparison with a geometric series of term x^2 , we have

$$\zeta(x) = \frac{2}{3}(1 + \frac{3}{5}x^2 + \frac{3}{7}x^4 + \dots) \leq \frac{2}{3}(1 + x^2 + x^4 + \dots) = \frac{2}{3}/(1-x^2).$$

It follows in particular that $0 \leq \zeta(x) \leq 1$ if $|x| \leq 3^{-1/2}$. With $x_j = (2j+\alpha+1)/(n+1)$ playing the rôle of x , we obtain

$$\beta_k = \beta_0 \exp \left(-2 \sum_{j=0}^{k-1} x_j - \sum_{j=0}^{k-1} \zeta(x_j) x_j^3 \right).$$

Keeping in mind the arithmetic series formula $1 + 2 + \dots + (k-1) = k(k-1)/2$, we have

$$\sum_{j=0}^{k-1} x_j = \frac{k(k-1)}{n+1} + \frac{(\alpha+1)k}{n+1} = \frac{k^2}{n} + \gamma_1 \frac{k}{n}$$

where $\gamma_1 = \gamma_1(k, n)$ is bounded between 0 and 1. For the remaining term, in view of the boundedness of $\zeta((2k+\alpha)/n)$ for all sufficiently large n , it will suffice to estimate the second sum by comparison with integrals,

$$0 \leq \sum_{j=0}^{k-1} \zeta(x_j) x_j^3 \leq \frac{1}{n^3} \sum_{j=0}^{k-1} (2j+2)^3 = \frac{8}{n^3} \sum_{j=1}^k j^3 \leq \frac{8}{n^3} \int_1^{k+1} x^3 dx \leq \frac{2(k+1)^4}{n^3}.$$

It follows that

$$\sum_{j=0}^{k-1} \zeta(x_j) x_j^3 = \gamma_2 \frac{k^4}{n^3}$$

where we may conservatively bound $\gamma_2 = \gamma_2(k, n)$ between 0 and 32. Stitching the terms together, we obtain

$$\beta_k = \beta_0 \exp \left(-\frac{2k^2}{n} - \gamma_1 \frac{k}{n} - \gamma_2 \frac{k^4}{n^3} \right).$$

If $k^4 \ll n^3$ then the terms k/n and k^4/n^3 will be subdominant compared to the term $2k^2/n$. Isolating the dominant term we hence obtain

$$\beta_k = \beta_0 e^{-2k^2/n} \left(1 - \gamma_3 \frac{k}{n} - \gamma_4 \frac{k^4}{n^3} \right)$$

where γ_3 and γ_4 depend on k and n but are bounded in absolute value. It follows that $\beta_k \sim \beta_0 e^{-2k^2/n}$ if $k^4/n^3 \rightarrow 0$. In view of our asymptotic estimate for β_0 , we obtain the

LOCAL LIMIT THEOREM Suppose $|k| \leq K_n$ where $K_n^4/n^3 \rightarrow 0$ as $n \rightarrow \infty$. Then

$$\beta_k \sim \frac{1}{\frac{1}{2}\sqrt{n}} \phi\left(\frac{k}{\frac{1}{2}\sqrt{n}}\right) \quad (n \rightarrow \infty)$$

uniformly in k . In other words, for each choice of $\epsilon > 0$ and every k in the range $-K_n \leq k \leq K_n$, we have

$$1 - \epsilon < \frac{\beta_k}{\frac{1}{2}\sqrt{n} \phi\left(\frac{k}{\frac{1}{2}\sqrt{n}}\right)} < 1 + \epsilon$$

eventually for all sufficiently large n .

The local limit theorem provides an alternative path to de Moivre's theorem. It is clear that for $|k| \leq K_n$,

$$\phi(x) \sim \frac{1}{\frac{1}{2}\sqrt{n}} \phi\left(\frac{x}{\frac{1}{2}\sqrt{n}}\right)$$

uniformly for all x in the interval $(k - \frac{1}{2})/\frac{1}{2}\sqrt{n} \leq x \leq (k + \frac{1}{2})/\frac{1}{2}\sqrt{n}$. It follows that

$$\beta_k \sim \int_{(2k-1)/\sqrt{n}}^{(2k+1)/\sqrt{n}} \phi(x) dx$$

uniformly for all $|k| \leq K_n$. Now $P\{a < S_n^* < b\} = \sum \beta_k$ where the sum is over all k satisfying $\frac{1}{2}a\sqrt{n} - \frac{1}{2} < k < \frac{1}{2}b\sqrt{n} + \frac{1}{2}$. This range is well within the purview of the local limit theorem and, accordingly,

$$P\{a < S_n^* < b\} \sim \sum_{k: a - \frac{1}{\sqrt{n}} < \frac{2k}{\sqrt{n}} < b + \frac{1}{\sqrt{n}}} \int_{(2k-1)/\sqrt{n}}^{(2k+1)/\sqrt{n}} \phi(x) dx.$$

The right-hand side differs from the Riemann integral $\int_a^b \phi(x) dx$ only by the addition of points at the boundary of integration and this results in a correction of no more than $2/\sqrt{n}$ which vanishes as $n \rightarrow \infty$. This establishes de Moivre's theorem.

7 Large deviations

It is common to read into de Moivre's theorem rather more than may legitimately be deduced. Couched in terms of the number S_n of successes in n tosses of a fair coin, the theorem says that

$$P\left\{ \frac{1}{2}n + \frac{1}{2}a\sqrt{n} < S_n < \frac{1}{2}n + \frac{1}{2}b\sqrt{n} \right\} \rightarrow \int_a^b \phi(x) dx$$

as $n \rightarrow \infty$. In other words, the theorem asserts that S_n exhibits normal tendency in a range of order \sqrt{n} on either side of $n/2$. It is not the case, however, that the distribution of S_n converges to the normal everywhere in its range, especially when the deviations from $n/2$ are of the order of n itself. This is the province of large deviations.

This comment notwithstanding, it is clear that the local limit theorem allows us to deduce somewhat more than advertised and we may allow $a = a_n$ and $b = b_n$ in de Moivre's theorem to depend on n provided we interpret the limiting result (5.1) properly: if $a_n \rightarrow \infty$ then both sides of (5.1) tend to zero and we replace the limit by the condition that the ratio of the two sides tends to one. It will suffice to consider one-sided results because of the following observation.

LEMMA 1 Suppose $a_n \rightarrow \infty$. Then

$$\frac{P\{S_n^* > a_n + \epsilon a_n\}}{P\{S_n^* > a_n\}} \rightarrow 0 \quad (n \rightarrow \infty)$$

for every choice of $\epsilon > 0$.

PROOF: With notation as before we have

$$\frac{P\{S_n^* > a_n + \epsilon a_n\}}{P\{S_n^* > a_n\}} = \frac{\sum_{j \geq 0} \beta_{s_n+j}}{\sum_{j \geq 0} \beta_{t_n+j}} \quad (7.1)$$

where s_n and t_n are integer values differing by at most one unit from $(1 + \epsilon)a_n\sqrt{n}/2$ and $a_n\sqrt{n}/2$, respectively. For the ratio of successive terms of the binomial we obtain from (6.8) that

$$\frac{\beta_{k+1}}{\beta_k} = \frac{1 - x_k}{1 + x_k} < 1 - x_k \leq e^{-x_k} < e^{-2k/n}$$

where the penultimate step follows by the elementary inequality $1 - x \leq e^{-x}$ and the final step follows because, for $k \leq n/2$,

$$x_k = \frac{2k + \alpha + 1}{n + 1} \geq \frac{2k + 1}{n + 1} \geq \frac{2k}{n}.$$

An easy induction now shows that

$$\frac{\beta_{s_n+j}}{\beta_{t_n+j}} < e^{-2(s_n-t_n)(t_n+j)/n} < e^{-\epsilon a_n^2/2}.$$

As the ratio of successive terms in the series in the numerator and denominator of (7.1) share the common exponential bound on the right, we see that

$$\frac{P\{S_n^* > a_n + \epsilon a_n\}}{P\{S_n^* > a_n\}} < e^{-\epsilon a_n^2/2} \rightarrow 0$$

as the exponent tends to $-\infty$ for any choice of $\epsilon > 0$. ▶

This result shows that $P\{a_n < S_n^* \leq a_n + \epsilon a_n\} \sim P\{S_n^* > a_n\}$ or, informally, that most of the mass in excess of a_n is concentrated in a region near a_n . A corresponding result also holds for the normal tail.

LEMMA 2 Suppose $a_n \rightarrow \infty$ as $n \rightarrow \infty$. Then

$$\int_{a_n + \epsilon a_n}^{\infty} \phi(x) dx / \int_{a_n}^{\infty} \phi(x) dx \rightarrow 0$$

for every fixed $\epsilon > 0$.

PROOF: It is clear that

$$\int_{a_n}^{\infty} \phi(x) dx \geq \int_{a_n}^{a_n + \epsilon a_n} \phi(x) dx \geq \frac{\epsilon a_n e^{-(1+\epsilon)^2 a_n^2 / 2}}{\sqrt{2\pi}}$$

as the integrand decreases monotonically. Coupled with the upper bound for the tail from Lemma 1.3, we obtain

$$\int_{a_n + \epsilon a_n}^{\infty} \phi(x) dx / \int_{a_n}^{\infty} \phi(x) dx \leq \frac{\sqrt{\pi}}{\sqrt{2} \epsilon a_n}$$

and the right-hand side tends to zero as $a_n \rightarrow \infty$. ▶

All is now in readiness for our large deviation result. It is clear from the discussion following the local limit theorem that

$$P\{a_n < S_n^* \leq a_n + \epsilon a_n\} \sim \int_{a_n}^{a_n + \epsilon a_n} \phi(x) dx.$$

The left-hand side is asymptotic, *vide* Lemma 1, to $P\{S_n^* > a_n\}$, while the right-hand side is asymptotic, *vide* Lemma 2, to $\int_{a_n}^{\infty} \phi(x) dx$, and we have proved the following

LARGE DEVIATION THEOREM Suppose the sequence $\{a_n\}$ grows unboundedly with n in such a way that $a_n n^{-1/4} \rightarrow 0$. Then

$$P\{S_n^* > a_n\} \sim \int_{a_n}^{\infty} \phi(x) dx \quad (n \rightarrow \infty).$$

For all its esoteric-looking nature the large deviation theorem is of frequent use in applications; the reader will find examples in rather different domains in the concluding sections of this chapter.

Thus, normal tendency persists well into the binomial tail; roughly speaking, the variable S_n is governed asymptotically by a normal law as long as we are concerned with deviations from $n/2$ up to order $n^{3/4}$. We will return fortified with more ammunition to the issue of even larger deviations of order n itself from $n/2$ in Chapter XVII.



8 The limits of wireless cohabitation

The binomial distribution crops up naturally in a variety of applications. Our first example is drawn from digital wireless communications; the author may perhaps be forgiven for choosing an illustration from his own work.

Suppose m users share a common radio channel in what is called a direct sequence, code division multiple access spread spectrum system for digital communications. To keep transmissions distinct, each user is allocated a random spreading sequence (or “signature”) which she uses to modulate each bit she transmits. I will provide here a sanitised view of how the system functions.

For our purposes it will be convenient to think of each bit as a -1 or a $+1$ (instead of the more conventional 0 and 1). For $1 \leq k \leq m$, user k is allocated a distinct spreading sequence $Z_k = (Z_{k1}, \dots, Z_{kn})$ where each Z_{kj} is -1 or $+1$. The “length” n of the sequence is a measure of the available bandwidth in the communication channel; for our purposes we only assume that it is large. For each bit $\beta_k \in \{-1, +1\}$ that user k wishes to communicate, she actually transmits the weighted spreading sequence $\beta_k Z_k = \beta_k (Z_{k1}, \dots, Z_{kn})$. The receiver gets a superposition of the transmissions of each of the users and generates the vector $R = \sum_{k=1}^m \beta_k Z_k$. The simplest strategy to recover the individual bits β_1, \dots, β_m from the mixture is to match the received vector $R = (R_1, \dots, R_n)$ with each of the individual spreading sequences in turn (the so-called *matched filter receiver*). For each $1 \leq \ell \leq m$, form the inner product

$$\langle R, Z_\ell \rangle = \sum_{j=1}^n R_j Z_{\ell j} = \sum_{j=1}^n \sum_{k=1}^m \beta_k Z_{kj} Z_{\ell j} = n \beta_\ell + \sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq \ell}}^m \beta_k Z_{kj} Z_{\ell j}.$$

If the sum on the right is less than n in absolute value then the inner product on the left will have the same sign as the transmitted bit β_ℓ and the estimate $\hat{\beta}_\ell = \text{sgn}\langle R, Z_\ell \rangle$ formed by taking the sign of the inner product will coincide with β_ℓ . Multiplying both sides by β_ℓ we obtain the following criterion: *the bit estimate $\hat{\beta}_\ell$ at the receiver will coincide with the transmitted bit β_ℓ if $\beta_\ell \langle R, Z_\ell \rangle$ is strictly positive or, equivalently, if $\sum_{j=1}^n \sum_{k \neq \ell} \beta_k \beta_\ell Z_{kj} Z_{\ell j} > -n$.*

Good statistical properties emerge if the bits in the spreading sequences are chosen randomly; accordingly, we assume that the double sequence $\{Z_{kj}\}$ is obtained by repeated independent trials of a chance experiment, each trial producing values -1 or $+1$ only, each with probability $1/2$. In this context it is natural to say that $\{Z_{kj}\}$ is a *sequence of independent random variables*. It is now not difficult to see that the doubly indexed sequence $X_{kj} = \beta_k \beta_\ell Z_{kj} Z_{\ell j}$ with $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, m\} \setminus \{\ell\}$ constitutes a sequence of $N = n(m-1)$ independent chance variables, each X_{kj} taking values -1 and $+1$ only with probability $1/2$ apiece. (Each X_{kj} contains the unique term Z_{kj} and the independence and symmetry of the Z_{kj} s carries the argument through.) It follows that $S_N = \sum_j \sum_{k \neq \ell} X_{kj}$ is a sum of N independent, ± 1 random variables. A trivial recentring and scaling shows now that S_N is governed by the binomial distribution. Indeed, $\frac{1}{2}(X_{kj} + 1)$ takes values 0 or 1 with probability $1/2$ apiece, so that, identifying 1 with “success” and 0 with “failure”, $\frac{1}{2}(S_N + N) = \sum_j \sum_{k \neq \ell} \frac{1}{2}(X_{kj} + 1)$ represents the number of successes in N tosses of a fair coin, and is hence governed by the distribution $b_N(i) = \binom{N}{i} 2^{-N}$. As the distribution of S_N is symmetric about 0 , the probability that the bit β_ℓ is *not* recovered at the receiver is hence given by

$$P\{\hat{\beta}_\ell \neq \beta_\ell\} = P\{S_N \leq -n\} = P\{S_N \geq n\} = \sum_{i \geq \frac{1}{2}(N+n)} b_N(i). \quad (8.1)$$

Suppose now that the number of users $m = m_n$ increases with n so that $mn^{-1/3} \rightarrow \infty$ or, equivalently, $nN^{-3/4} \rightarrow 0$, as $n \rightarrow \infty$. The large deviation theorem of Section 7 then tells us that the binomial sum on the right in (8.1) is asymptotically

equivalent (in the sense that the ratio of the two terms tends to one) to the normal tail integral $\int_{n/\sqrt{N}}^{\infty} \phi(x) dx$ which, by Lemma 1.3, is bounded above by $\frac{1}{2}e^{-n^2/(2N)}$. (Theorem X.1.1 provides a sharper estimate.) Boole's inequality allows us to wrap up: for all sufficiently large n , the probability that one or more of the transmitted bits β_1, \dots, β_m is not recovered is bounded by

$$\mathbf{P}\left(\bigcup_{\ell=1}^m \{\hat{\beta}_\ell \neq \beta_\ell\}\right) \leq \sum_{\ell=1}^m \mathbf{P}\{\hat{\beta}_\ell \neq \beta_\ell\} \leq \frac{m}{2} \exp\left(-\frac{n^2}{2N}\right) \leq \frac{m}{2} \exp\left(-\frac{n}{2m}\right),$$

eventually. If $m = m_n \leq n/(2 \log n)$ the right-hand side is bounded above by $1/(4 \log n)$ and hence tends to zero. Thus, the channel can *simultaneously* accommodate at least $n/(2 \log n)$ users with a guarantee of successful recovery of transmitted bits for all users. A more refined analysis shows that there is actually a *phase transition* that occurs around this value. The *theorem*: *Suppose $\epsilon > 0$ is fixed. If $m_n \leq n/(2 \log n)$ then the probability that all transmitted bits are recovered without error tends to one; if, on the other hand, $m_n \geq (1 + \epsilon)n/(2 \log n)$ then the probability that all transmitted bits are recovered without error tends to zero.*²



9 When memory fails

Our next illustration of the binomial in action, with a similar framework to that of the previous section—but a very different setting—is drawn from computational neuroscience.

In 1943, W. S. McCulloch and W. H. Pitts proposed their famous model of a biological neuron as a computational element that forms the sign of a linear form of its inputs.³ I shall describe in this section the rudiments of an influential model for associative memory popularised by J. J. Hopfield⁴ in which information is stored in an interconnected network of n McCulloch–Pitts elements whose outputs at epoch t are fed back to constitute their inputs at the next epoch $t + 1$.

I shall refer to the linear threshold elements of McCulloch and Pitts as *neurons* though the reader will bear in mind that these stylised mathematical neurons are a far cry from the biological complexity of the real thing. For each $1 \leq i \leq n$, the i th neuron in the network is characterised by a system of real *weights* w_{i1}, \dots, w_{in} . At each epoch t , the current neural outputs $z_1(t), \dots, z_n(t) \in \{-1, +1\}$ are fed back as inputs and, for $1 \leq i \leq n$, the i th neuron updates its output at epoch $t + 1$ according to the sign of the linear form $\sum_{j=1}^n w_{ij} z_j(t)$, that is, $z_i(t + 1) = \text{sgn} \sum_{j=1}^n w_{ij} z_j(t)$. (The specifics of the ordering of updates makes little matter for our purposes but, for definiteness, we

²For refinements and extensions, see S. S. Venkatesh, “CDMA capacity”, *Proceedings of the Conference on Information Sciences and Systems*, Princeton University, March 15–17, 2000.

³W. S. McCulloch and W. H. Pitts, “A logical calculus of the ideas immanent in nervous activity”, *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

⁴J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities”, *Proceedings of the National Academy of Sciences of the USA*, vol. 79, no. 8, pp. 2554–2558, 1982.

may suppose that all neurons update their outputs in synchrony.) The neural outputs at time t hence define a *temporal state* $\mathbf{z}(t) = (z_1(t), \dots, z_n(t))$ of the network and the sequence of state updates, $\mathbf{z}(1) \mapsto \mathbf{z}(2) \mapsto \dots \mapsto \mathbf{z}(t) \mapsto \mathbf{z}(t+1) \mapsto \dots$, then defines a dynamics on the vertices $\{-1, +1\}^n$ of the cube in n dimensions. Of particular interest are states $\mathbf{z} = (z_1, \dots, z_n)$ that are fixed points of the network, i.e., states that satisfy the relation $z_i = \text{sgn} \sum_{j=1}^n w_{ij} z_j$, or equivalently, $\sum_{j=1}^n w_{ij} z_i z_j > 0$, for each $i = 1, \dots, n$. Such states are stable in the system dynamics and may be thought of as embedded system memory—if a state trajectory leads into a fixed point at some epoch then the vector of neural outputs that results will remain immutable from that point onwards. The fixed points of the network are determined completely by the matrix of weights $[w_{ij}]$ characterising the system. It is of interest in such a setting to ask whether it is possible to select a system of weights $[w_{ij}]$ so that a *given* set of vertices of the cube representing desired memories can be made fixed points of the network.

Suppose the targeted vertices $\mathbf{Z}^{(k)} = (Z_1^{(k)}, \dots, Z_n^{(k)})$ with k ranging from 1 to m are a collection of m random vertices of the cube where, as in the setting of Section 8, the mn variables $Z_j^{(k)}$ represent the results of repeated independent trials, each trial producing values -1 or $+1$ only, each with probability $1/2$. For each i and j , the simplest reinforcement model for weight formation now suggests setting

$$w_{ij} = \sum_{k=1}^m Z_i^{(k)} Z_j^{(k)}. \quad (9.1)$$

We should immediately observe that $w_{ij} = w_{ji}$ so that the interaction weights are symmetric. The motivation is as follows: for each targeted state $\mathbf{Z}^{(k)}$, the desired outputs $Z_i^{(k)}$ and $Z_j^{(k)}$ of neurons i and j mutually reinforce each other in the weight of the symmetric connection between them. This is the persuasive model of neural learning proposed by D. O. Hebb.⁵ The resulting matrix of weights for the network essentially maintains a superposed template of each of the vertices $\mathbf{Z}^{(k)}$ and intuition may lead the reader to feel that the vertices should be fixed points of the network if m is not too large.

For any given $1 \leq \ell \leq m$, the vertex $\mathbf{Z}^{(\ell)}$ is a fixed point if, for each i ,

$$0 < \sum_{j=1}^n w_{ij} Z_i^{(\ell)} Z_j^{(\ell)} = \sum_{j=1}^n \sum_{k=1}^m Z_i^{(k)} Z_j^{(k)} Z_i^{(\ell)} Z_j^{(\ell)} = n + \sum_{j=1}^n \sum_{\substack{k=1 \\ k \neq \ell}}^m Z_i^{(k)} Z_j^{(k)} Z_i^{(\ell)} Z_j^{(\ell)}. \quad (9.2)$$

For fixed i and ℓ , the $N = n(m - 1)$ summands $X_j^{(k)} = Z_i^{(k)} Z_j^{(k)} Z_i^{(\ell)} Z_j^{(\ell)}$ with $j \in \{1, \dots, n\}$ and $k \in \{1, \dots, m\} \setminus \{\ell\}$ form a system of independent random variables, each taking values -1 or $+1$ only, each again with probability $1/2$. (As in Section 8, the presence of the unique term $Z_j^{(k)}$ in each summand is decisive.) It follows that we may write the inequality (9.2) in the form $S_N > -n$ where $S_N = \sum_{j=1}^n \sum_{k \neq \ell} X_j^{(k)}$ is the sum of N independent, ± 1 -valued random variables. In this notation, the probability that the inequality (9.2) is violated is hence given by $P\{S_N \leq -n\}$. Arguing as in the previous section, the trivial recentering and scaling $\frac{1}{2}(S_N + N) = \sum_j \sum_{k \neq \ell} \frac{1}{2}(X_j^{(k)} + 1)$ represents the accumulated number of successes in N tosses of a fair coin, and hence

⁵D. O. Hebb, *The Organization of Behavior*. New York: John Wiley, 1949.

has the binomial distribution $b_N(\cdot)$. Proceeding with the analysis as before, the large deviation theorem of Section 7 shows that $P\{S_N \leq -n\} = P\{S_N \geq n\} \leq \frac{1}{2}e^{-n/(2m)}$, eventually, provided $m = m_n$ grows with n so that $mn^{-1/3} \rightarrow \infty$. By Boole's inequality, the probability that at least one of the mn inequalities (9.2) is violated is hence less than $\frac{1}{2}mne^{-n/(2m)}$ for large enough n . If $m = m_n \leq n/(4 \log n)$ this probability is bounded above by $1/(8 \log n)$ and hence vanishes asymptotically. It follows that almost all collections of $n/(4 \log n)$ vertices or fewer can be made fixed points of a network with interconnection weights specified by the Hebbian learning algorithm (9.1). A more searching examination actually shows that a phase transition emerges. The *theorem*: Suppose $\epsilon > 0$. If $m_n \leq n/(4 \log n)$ then the probability that all the selected m_n vertices are fixed points tends to one; if, on the other hand, $m_n \geq (1 + \epsilon)n/(4 \log n)$, with probability approaching one at least one of the selected vertices will not be a fixed point.⁶ The resemblance with the theorem of the previous section is striking.

10 Problems

1. The gamma function. The function $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$, defined for all $t > 0$, plays a particularly important rôle in analysis and is called the gamma function. Evaluate $\Gamma(1)$ and $\Gamma(1/2)$. (The latter integral is made transparent by making the change of variable $x = y^2/2$ inside the integral.)

2. Another proof of de Moivre's theorem. Stirling's formula for the factorial says that $n!/\left[\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\right] \rightarrow 1$ as $n \rightarrow \infty$. (The reader who does not know this result will find two elementary proofs from rather different viewpoints in Section XIV.7 and again in Section XVI.6.) By repeated applications show that $\lim_{n \rightarrow \infty} \binom{n}{\lceil n/2 \rceil} 2^{-n} / \sqrt{\frac{2}{\pi n}} = 1$. Deduce the local limit theorem of Section 6 and thence de Moivre's theorem.

3. Another derivation of the law of large numbers. Derive the weak law of large numbers for coin tosses (V.6.3) from de Moivre's theorem.

4. Stability of the normal density. Writing $\sigma^2 = \sigma_1^2 + \sigma_2^2$, use the local limit theorem of Section 6 to conclude that $\frac{1}{\sigma} \phi\left(\frac{x}{\sigma}\right) = \int_{-\infty}^{\infty} \frac{1}{\sigma_1} \phi\left(\frac{x-t}{\sigma_1}\right) \cdot \frac{1}{\sigma_2} \phi\left(\frac{t}{\sigma_2}\right) dt$. Only the statement of the local limit theorem is needed for this problem.

5. A grossly simplified model of portfolio growth. Suppose $0 < \alpha < 1 < \beta$. In a market uptick wealth increases by β and in a market downtick it decreases by α . Suppose our portfolio is initially worth one unit, $W_0 = 1$, wealth getting adjusted upward or downward at each epoch. Immediately after the n th market epoch, our wealth is given by $W_n = \alpha W_{n-1}$ if there was a downtick at epoch n and by $W_n = \beta W_{n-1}$ if there was an uptick. Suppose that at each epoch upticks and downticks are governed by the toss of a fair coin. Show that $\lim_n P\{W_n > (\alpha\beta)^{n/2}(\beta/\alpha)^{t\sqrt{n}/2}\} = \int_t^\infty \phi(x) dx$ for every t . Hence conclude that wealth grows exponentially if $\beta > 1/\alpha$ and decreases exponentially if $\beta < 1/\alpha$. What happens if $\beta = 1/\alpha$? [Hint: Write $W_n = W_{n-1} \alpha^{1-X_n} \beta^{X_n}$, where $X_n = 0$ or 1 only, each with probability $1/2$, and take logarithms.]

⁶R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the Hopfield associative memory", *IEEE Transactions on Information Theory*, vol. IT-33, pp. 461–482, 1987.

6. *Stability of the Poisson distribution.* Suppose X_1, X_2, \dots are the outcomes of repeated independent trials, each X_i taking positive integer values according to the distribution $p(k) = P\{X_i = k\} = e^{-1}/k!$ for $k \geq 0$. Form the partial sums $S_n = X_1 + \dots + X_n$. Determine the distribution of S_n for each n . [Hint: Induction.]

7. *Continuation, a central limit theorem for the Poisson distribution.* Fix $a < b$ and, for each positive t , let \mathbb{A}_t be the collection of positive integers k satisfying $t + a\sqrt{t} < k < t + b\sqrt{t}$. Use Markov's method to prove Laplace's formula $\lim_{t \rightarrow \infty} \sum_{k \in \mathbb{A}_t} e^{-t - \frac{t^k}{k!}} = \int_a^b \phi(x) dx$. [Hint: Begin by allowing t to tend to infinity through the positive integers n . With S_n as in the previous problem, consider the normalised variable $S_n^* = (S_n - n)/\sqrt{n}$ and write down a sum for $P\{a < S_n^* < b\}$. Now consider a decomposition of the summation index k in the form $k = k_1 + \dots + k_n$ and utilise Markov's method by exploiting the fact that the trials X_1, \dots, X_n are independent.]

8. *Weyl's equidistribution theorem.* For any real x , let $(x) = x \bmod 1 = x - [x]$ denote the fractional part of x . (In standard notation, $[x]$ represents the integer part of x ; our notation for the fractional part is not standard.) As n steps through the positive integers, the values $x_n = (nx)$ then determine a sequence of points in the unit interval. In 1917, H. Weyl proved that if x is irrational then the sequence $\{x_n\}$ is equidistributed in $(0, 1)$. In other words, fix any $0 \leq a < b \leq 1$ and let $\xi_n = 1$ if $a < x_n < b$ and let $\xi_n = 0$ otherwise. For each n , let $v_n = (\xi_1 + \dots + \xi_n)/n$ denote the fraction of the numbers x_1, \dots, x_n which fall in (a, b) . Then $v_n \rightarrow b - a$ as $n \rightarrow \infty$. Prove Weyl's theorem by introducing the periodic function $f(x)$ of period 1, with the form given by (4.1) in the interval $0 \leq x < 1$, and using Fourier series instead of Fourier integrals.

The succeeding problems build upon Problems V.12–15.

9. *Bent coins.* In repeated tosses of a bent coin with success probability p (and corresponding failure probability $q = 1 - p$) let S_n denote the number of successes after n tosses and let $b_n(k; p)$ denote the probability that $S_n = k$. Starting with the expression derived in Problem V.14 show that $b_n(k; p) = \binom{n}{k} p^k q^{n-k}$.

10. *The central term of the binomial.* With notation as in the previous problem, for each fixed n and p , show that there is a unique integer m such that $b_n(k; p) \geq b_n(k-1; p)$ for all $k \leq m$ and show that $(n+1)p - 1 < m \leq (n+1)p$. Under what condition is it true that $b_n(m; p) = b_n(m-1; p)$?

11. *Laplace's theorem.* With S_n as in Problem 9, denote the normalised number of successes by $S_n^* = (S_n - np)/\sqrt{npq}$. For any $a < b$, using Markov's method show that $P\{a < S_n^* < b\} \rightarrow \int_a^b \phi(x) dx$ asymptotically as $n \rightarrow \infty$. This extension of de Moivre's theorem to bent coins is due to Laplace.

12. *A local limit theorem for bent coins.* Let $\beta_k = b_n(m+k; p)$ with m as in Problem 10. Suppose $\{K_n\}$ is any sequence of positive integers satisfying $K_n/(npq)^{2/3} \rightarrow 0$ as $n \rightarrow \infty$. (Tacitly, p , hence also q , are allowed to depend on n .) Show that the asymptotic estimate $\beta_k \sim \frac{1}{\sqrt{npq}} \phi\left(\frac{k}{\sqrt{npq}}\right)$ holds uniformly as $n \rightarrow \infty$ for all k in the interval $-K_n \leq k \leq K_n$. [Note that the range of validity is reduced when the coin is unfair; symmetry has its advantages.]

13. *A large deviation theorem for bent coins.* Suppose the sequence $\{a_n\}$ grows in such a way that $a_n \sqrt{pq} \rightarrow \infty$ and $a_n/(npq)^{1/6} \rightarrow 0$. Then $P\{S_n^* > a_n\} \sim \int_{a_n}^{\infty} \phi(x) dx$.

VII

Probabilities on the Real Line

Chance experiments whose outcomes are real-valued play an important rôle in shaping our probabilistic intuition and it is fortunate that the structure of the real line lends itself to particular simplicities in specifying probability measures. The large scope of applications of this type of structure prompts us to consider this setting in more detail.

C 1–6, 9, 10
A 7, 8

The outcome of a probability experiment is called a *random variable* when the sample space consists of a set of real numbers. A generic random variable will be denoted by X and connotes the random real outcome of the experiment; at need letters like Y , Z , etc., will be pressed into service to denote other random variables, the typical use of capitals in the notation serving to identify random variables from “free” algebraic variables like x , y , and z .

We will focus in this chapter on the two simplest variations—when the distributions corresponding to the random variables are either arithmetic or continuous—leaving the development of the general theory to Chapter XII.

1 Arithmetic distributions

Probability experiments with discrete outcomes shape much of our common intuition and feel for probability as a subject. Even in apparently simple settings, however, there are pitfalls for the unwary and unexpected sophistications. We begin with this setting first.

In many natural situations such as the roll of a die, the count of the number of eggs laid by a crustacean, or the number of radioactive particles striking a geiger counter, the possible outcomes of the experiment may be identified with a *regularly spaced* set of numbers, either finite or denumerably infinite, which constitutes a discrete sample space. We will focus first on this setting with a view to delineating the features of the most important of the discrete distributions—the so-called arithmetic distributions.

A random variable X that can only take values $0, \pm a, \pm 2a, \dots$ for some fixed a is said to be *arithmetic*, the nomenclature used also to describe

its distribution.¹ It is also permissible to call an arithmetic random variable a *lattice* random variable. The largest a for which an arithmetic random variable takes values in the set $\{ak, \text{integer } k\}$ is called its *span*.

The span represents a scale factor and does not play a significant rôle for our purposes in this chapter. *We henceforth consider arithmetic random variables with span 1, that is to say, random experiments whose outcomes take only integer values $0, \pm 1, \pm 2, \dots$*

We begin accordingly with an arithmetic random variable X taking integer values only. The setting is that of Section I.7 and for convenience we summarise the notation that was introduced there.

Any subset of integers \mathbb{A} constitutes an event in this setting. We adopt the flexible and graphic notation $\{X \in \mathbb{A}\}$ to denote the event that the experimental outcome X takes value in \mathbb{A} ; in particular, the singleton event $\{X = k\}$ occurs if the outcome of the experiment is the fixed integer k . This notation has the advantage of not only identifying the set of sample points \mathbb{A} that constitutes the event at hand but also the underlying experiment through the specification of X ; this flexibility in notation will prove to be useful when several experiments are being considered concurrently with outcomes, say, X, Y, Z , etc., and it becomes necessary to identify which experiment is currently under discussion.

The probability that the arithmetic random variable X takes the fixed value k specifies a function defined on the integers,

$$\mathbf{P}\{X = k\} = p(k) \quad (\text{integer } k),$$

called the (*probability*) *distribution* of X .² Of course, it is clear that the distribution $\{p(k)\}$ satisfies $p(k) \geq 0$ and $\sum_k p(k) = 1$. The distribution of X completes the specification of the probability space; if \mathbb{A} is any collection of integers, the probability that X takes value in \mathbb{A} is given by $\mathbf{P}\{X \in \mathbb{A}\} = \sum_{k:k \in \mathbb{A}} p(k)$, the sum, as indicated, running over all integers k in \mathbb{A} .

Positive arithmetic random variables X play an important rôle. These are random variables whose distribution $p(k)$ is identically zero for $k < 0$.

Random experiments whose outcomes range only over a *finite* collection of integers are covered in this setting by the simple expedient of setting $p(k)$ to zero if the integer k is not a possible outcome of the experiment. This convention allows us to consider arithmetic variables in a unified framework without constantly making annoying distinctions between experiments with a finite number of possible integer-valued outcomes and experiments with a denumerably infinite number of possible integer-valued outcomes. Sums without explicitly specified limits should then be assumed to range over all integers. *We achieve some additional compactness in presentation by adopting the convention that,*

¹The terminology is not completely standard; some authors define an arithmetic random variable as one that takes values in a set $\{c + ak, k \text{ integer}\}$.

²For a discrete random variable the distribution is also called the *probability mass function*.

in any given example or application, the distribution is to be assumed to be identically zero wherever its value is not explicitly specified.

The mean (or expectation) of an arithmetic random variable X with distribution $\{p(k)\}$ (or, more compactly, the mean of X or the mean of the distribution p) is defined by

$$E(X) = \sum_k kp(k)$$

provided the sum converges absolutely. The mean of X (if it exists) is a weighted average of the values that X can assume and is a natural extension of the notion of the arithmetic mean of a finite set of numbers. For our present purposes, the mean represents a crude sort of single-number approximation of X . The notion has far-reaching analytical consequences, however, and we will explore these in later chapters.

Other measures can be constructed to add more information about X ; of these the most important is the variance. Suppose X has mean μ . The variance of X is defined by

$$\text{Var}(X) = \sum_k (k - \mu)^2 p(k)$$

if the sum is convergent. Expanding the square in the summand allows us to write the variance in the equivalent alternative form

$$\text{Var}(X) = \sum_k k^2 p(k) - 2\mu \sum_k kp(k) + \mu^2 \sum_k p(k) = \sum_k k^2 p(k) - \mu^2$$

which is convenient for calculation. The standard deviation of X is the (positive) square-root of the variance. It is conventional to write σ^2 for the variance; the standard deviation is then σ . The variance intuitively captures the extent of the (squared) spread of possible values of X around its mean μ . If the variance is small then the likely values of X are concentrated around the mean; conversely, if the variance is large then it is not unlikely that X takes value far from its mean.

EXAMPLES: 1) *The roll of a die.* If X represents the outcome of the roll of a fair die then the values 1 through 6 are equally likely and its distribution is specified by $p(1) = \dots = p(6) = 1/6$ (with all other $p(k)$ identically zero). The mean of X is

$$\mu = \frac{1}{6}(1 + 2 + \dots + 6) = 3.5,$$

and its variance is

$$\sigma^2 = \frac{1}{6}(1^2 + 2^2 + \dots + 6^2) - (3.5)^2 = 2.916\dots.$$

The corresponding standard deviation is $\sigma = 1.707\dots$

If the die is n -sided then $p(k) = 1/n$ for $k = 1, \dots, n$ and is zero otherwise. The mean and variance are now given by

$$\mu = \sum_{k=1}^n \frac{1}{n} k = \frac{1}{2}(n+1), \quad \sigma^2 = \sum_{k=1}^n \frac{1}{n} k^2 - \frac{1}{4}(n+1)^2 = \frac{1}{12}(n^2 - 1).$$

2) A divergent sum. Consider an arithmetic random variable taking values in the positive integers with distribution $p(k) = 1/(k(k+1))$ for $k \geq 1$. It is easy to verify by partial fractions that this defines a proper distribution as

$$\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \sum_{k=1}^{\infty} \left(\frac{1}{k} - \frac{1}{k+1} \right) = \left(1 - \frac{1}{2} \right) + \left(\frac{1}{2} - \frac{1}{3} \right) + \left(\frac{1}{3} - \frac{1}{4} \right) + \dots$$

and the sum telescopes to the value 1. On the other hand, the reader is well aware that the harmonic series diverges so that $\sum_{k=1}^{\infty} kp(k) = \sum_{k=1}^{\infty} (k+1)^{-1}$ diverges to $+\infty$. It follows that the mean of this distribution is $+\infty$. If instead $p(k) = 1/(2|k|(|k|+1))$ for $|k| \geq 1$ with $p(0) = 0$, then $\sum_k kp(k)$ is not absolutely convergent and the mean does not exist. ►

2 Lattice distributions

The considerations for a single arithmetic random variable extend quite naturally to probability experiments that result in two or more arithmetic variables. For instance, the outcome of an experiment consisting of the tossing of n dice is an n -tuple of numbers (X_1, \dots, X_n) each taking integer values from 1 through 6. We begin by considering random pairs first to settle the concepts.

A pair of arithmetic variables (X_1, X_2) resulting from the performance of a probability experiment is called a *lattice variable*. The sample space now corresponds to the set of *lattice points* (k_1, k_2) with k_1 and k_2 varying over the integers. Events now correspond to sets of integer pairs; if \mathbb{A} is any collection of lattice points (k_1, k_2) then $\{(X_1, X_2) \in \mathbb{A}\}$ denotes the event that the lattice variable (X_1, X_2) takes value in \mathbb{A} ; the singleton event $\{X_1 = k_1, X_2 = k_2\} = \{(X_1, X_2) = (k_1, k_2)\}$ corresponds to the event that X_1 takes value k_1 and X_2 takes value k_2 . The distribution of the lattice variable (X_1, X_2) is a positive function defined on integer pairs,

$$p(k_1, k_2) = P\{X_1 = k_1, X_2 = k_2\} \quad (\text{integer } k_1, k_2),$$

and normalised so that $\sum_{k_1, k_2} p(k_1, k_2) = 1$. Probabilities are now assigned in the usual fashion: $P\{(X_1, X_2) \in \mathbb{A}\} = \sum_{(k_1, k_2) \in \mathbb{A}} p(k_1, k_2)$ for any set of lattice points \mathbb{A} .

The distribution $p(k_1, k_2)$ of the pair (X_1, X_2) implicitly contains all relevant information about X_1 and X_2 individually. Indeed, additivity of probability measure says that

$$p_1(k_1) := P\{X_1 = k_1\} = P\{X_1 = k_1, -\infty < X_2 < \infty\} = \sum_{k_2} p(k_1, k_2),$$

$$p_2(k_2) := P\{X_2 = k_2\} = P\{-\infty < X_1 < \infty, X_2 = k_2\} = \sum_{k_1} p(k_1, k_2).$$

The functions $p_1(k_1)$ and $p_2(k_2)$ are called the *marginal* distributions of X_1 and X_2 , respectively. The mean and variance of X_1 and X_2 may now be computed in the usual fashion.

EXAMPLE 1) Three dice. In an experiment consisting of the roll of three dice suppose that all outcomes are equally likely. Let X_1 and X_2 represent the number of 1s and 6s that result, respectively. It is simplest to specify the distribution $p(k_1, k_2)$ in tabular form, rows representing k_1 and columns representing k_2 .

	0	1	2	3
0	$\frac{64}{216}$	$\frac{48}{216}$	$\frac{12}{216}$	$\frac{1}{216}$
1	$\frac{48}{216}$	$\frac{24}{216}$	$\frac{3}{216}$	0
2	$\frac{12}{216}$	$\frac{3}{216}$	0	0
3	$\frac{1}{216}$	0	0	0

Summing each row yields the distribution of X_1 while summing each column yields the distribution of X_2 ; by symmetry, of course, they will have the same distribution shown below.

k	0	1	2	3
$p_{1(2)}(k)$	$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$

Thus X_1 and X_2 have common mean $\frac{1}{216}(1 \cdot 75 + 2 \cdot 15 + 3) = \frac{1}{2}$ which is intuitive and satisfactory. The common variance is $\frac{1}{216}(1 \cdot 75 + 4 \cdot 15 + 9) - (\frac{1}{2})^2 = \frac{5}{12}$. ▶

New random variables can now be defined as a function of X_1 and X_2 . These derived variables will inherit their distribution from the (joint) distribution of X_1 and X_2 .

EXAMPLE 2) The first throw in craps, reprise. The dice game of craps begins with the roll of two dice. The gambler loses on the first throw if the sum of the face values is 2, 3, or 12 and wins on the first throw if the sum of the face values is 7 or 11. Let X_1 and X_2 represent the face values of the two dice. Assuming all outcomes are equally likely, each of the events $\{X_1 = k_1, X_2 = k_2\}$ with $1 \leq k_1, k_2 \leq 6$ has equal probability $1/36$. We may now consider the sum of the face values, $S = X_1 + X_2$, to be a new variable derived from the coordinate variables X_1 and X_2 . The distribution $q(k)$ of S was determined in Example I.2.2 from these considerations. The variable S does not have any “monotonicity” implications for the progress of the game—larger values are not necessarily better—and in consequence the mean and variance do not mean very much. But a formal calculation using the values for $q(k)$ from Table 1 in Example I.2.2 shows that the mean of S is $\sum_k k q(k) = 7$ (as symmetry would dictate) and its variance is $\sum_k (k - 7)^2 q(k) = 5.83 \dots$ ▶

The generalisation of these ideas to experiments resulting in several arithmetic variables is straightforward. A *lattice variable in n dimensions* is an n -tuple of arithmetic random variables (X_1, \dots, X_n) . Its distribution is a positive function of n -tuples of integers (k_1, \dots, k_n) (lattice points in n dimensions),

$$p(k_1, \dots, k_n) = P\{X_1 = k_1, \dots, X_n = k_n\} \quad (\text{integer } k_1, \dots, k_n),$$

normalised so that $\sum_{k_1, \dots, k_n} p(k_1, \dots, k_n) = 1$. For each i , the marginal distribution of the individual arithmetic variable X_i is obtained by “summing out” the variables k_j ($j \neq i$) from the (joint) distribution:

$$p_i(k_i) = P\{X_i = k_i\} = \sum_{k_j (j \neq i)} p(k_1, \dots, k_n).$$

Of course, this is just another manifestation of additivity.

The simplest lattice variables arise in the case of independent trials. The reader should recall that, at its most basic level, independence is a rule of multiplication of probabilities. In a lattice setting this principle takes a very simple form. Let us begin with random pairs to settle the intuition.

Suppose (X_1, X_2) is a lattice variable with distribution $p(k_1, k_2)$. We say that X_1 and X_2 are *independent random variables if there exist two distributions $p_1(k_1)$ and $p_2(k_2)$ such that $p(k_1, k_2) = p_1(k_1)p_2(k_2)$* . In this case it is not at all difficult to see that $p_1(k_1)$ and $p_2(k_2)$ must in fact be the marginal distributions of X_1 and X_2 , respectively. Indeed, by additivity,

$$P\{X_1 = k_1\} = \sum_{k_2} p(k_1, k_2) = p_1(k_1) \sum_{k_2} p_2(k_2) = p_1(k_1)$$

as the normalisation $\sum_{k_2} p_2(k_2) = 1$ is forced because p_2 is a distribution. It follows that $p_1(k_1)$ is indeed the marginal distribution of X_1 . An entirely similar argument shows that $p_2(k_2)$ is the marginal distribution of X_2 .

It is apparent that this setting is just that of independent trials where the successive outcomes X_1 and X_2 of the experiments are integers. It follows that events depending only on the outcome X_1 are independent of events depending only on the outcome X_2 . Or, what says the same thing in notation, if \mathbb{A}_1 and \mathbb{A}_2 are two collections of integers then

$$P\{X_1 \in \mathbb{A}_1, X_2 \in \mathbb{A}_2\} = P\{X_1 \in \mathbb{A}_1\} P\{X_2 \in \mathbb{A}_2\}.$$

For a quick check of the concept, the variables X_1 and X_2 in Example 1 are dependent as they fail the product test; the variables X_1 and X_2 in Example 2 are independent.

Extensions of the principle are straightforward. Suppose (X_1, \dots, X_n) has distribution $p(k_1, \dots, k_n)$. Then X_1, \dots, X_n are *independent lattice random variables if, and only if, the n-dimensional distribution $p(k_1, \dots, k_n)$ factors into a*

product of n one-dimensional distributions $p_1(k_1), \dots, p_n(k_n)$. In this case we may identify each $p_i(k_i)$ as the marginal distribution of the corresponding lattice variable X_i . As for the two-dimensional case, it is now simple to verify that X_1, \dots, X_n are independent if, and only if, the events $\{X_1 \in A_1\}, \dots, \{X_n \in A_n\}$ are independent for all choices of A_1, \dots, A_n .

CONVOLUTION SUMS

Sums of random variables are endemic in probability. The distribution of a sum is in general complicated but admits of a compact representation in the special and important case when the summands are independent.

Suppose X_1, \dots, X_n are independent arithmetic random variables with marginal distributions $p_1(k_1), \dots, p_n(k_n)$, respectively. Let $S_n = X_1 + \dots + X_n$. Our goal is to determine the (marginal) distribution $q_n(k)$ of the sum random variable S_n .

While the possible combinations of values of X_1, \dots, X_n that result in a given value of S_n appear dauntingly complex at first blush, an examination of the simplest sum $S_2 = X_1 + X_2$ lays bare the basic mechanism. For any given integer k , the event $\{S_2 = k\}$ occurs if, and only if, the events $\{X_1 = j\}$ and $\{X_2 = k - j\}$ occur jointly for some choice of integer j . By additivity, it follows that

$$\mathbf{P}\{S_2 = k\} = \sum_j \mathbf{P}\{X_1 = j, X_2 = k - j\} = \sum_j \mathbf{P}\{X_1 = j\} \mathbf{P}\{X_2 = k - j\},$$

the final step following as X_1 and X_2 are assumed independent. Of course, we could equally well have interchanged the rôles of X_1 and X_2 in the above argument. We hence obtain the two equivalent representations

$$q_2(k) = \sum_j p_1(j)p_2(k - j) = \sum_j p_1(k - j)p_2(j) \quad (2.1)$$

for the distribution of S_2 . Either sum on the right is called the *convolution* of the distributions p_1 and p_2 and denoted $(p_1 * p_2)(k)$. As seen above, the convolution operation $*$ is commutative and the convolution sum may be performed in any order.

EXAMPLES: 3) *Two dice, reprise.* In Example 2, the probabilities $q(k)$ for the sum of the face values of two thrown dice were inferred by direct counting. Alternatively, we may write $q(k) = (p * p)(k)$ where $p(k)$ is the fair die distribution placing equal mass on each of the points $1, \dots, 6$. As the arguments of $p = p_1 = p_2$ in the convolution (2.1) are hence constrained to be in the range from 1 to 6, it follows that

$$q(k) = \sum_{j=\max\{1, k-6\}}^{\min\{6, k-1\}} \frac{1}{36} = \begin{cases} \frac{1}{36}(k-1) & \text{if } 2 \leq k \leq 7, \\ \frac{1}{36}(13-k) & \text{if } 7 \leq k \leq 12. \end{cases}$$

4) *Generalisation.* The sum of the face values of two n -sided dice has distribution

$$q(k) = \frac{\min\{n, k-1\} - \max\{1, k-n\}}{n^2} = \begin{cases} \frac{1}{n^2}(k-1) & \text{if } 2 \leq k \leq n+1, \\ \frac{1}{n^2}(2n+1-k) & \text{if } n+1 \leq k \leq 2n. \end{cases}$$

The argument follows that for the special case $n = 6$ above. ►

Let us now consider the arithmetic random variables $S_{n-1} = X_1 + \dots + X_{n-1}$ and X_n . It is intuitively clear that S_{n-1} and X_n are independent and it is not hard to furnish a proof. For any integer k let \mathbb{A}_k be the set of $(n-1)$ -tuples (k_1, \dots, k_{n-1}) satisfying $k_1 + \dots + k_{n-1} = k$. For any given k and l then, it follows by additivity that

$$\begin{aligned} \mathbf{P}\{S_{n-1} = k, X_n = l\} &= \sum_{(k_1, \dots, k_{n-1}) \in \mathbb{A}_k} p(k_1, \dots, k_{n-1}, l) \\ &= \sum_{(k_1, \dots, k_{n-1}) \in \mathbb{A}_k} p_1(k_1) \cdots p_{n-1}(k_{n-1}) p_n(l) = q_{n-1}(k) p_n(l) \end{aligned}$$

as $q_{n-1}(k) = \sum_{(k_1, \dots, k_{n-1}) \in \mathbb{A}_k} p_1(k_1) \cdots p_{n-1}(k_{n-1})$ as a matter of definition. It follows that S_{n-1} and X_n are independent random variables. Now $S_n = S_{n-1} + X_n$ so that the problem is reduced to the problem of a sum of two independent random variables. Accordingly, we obtain the recursive specification $q_n(k) = (q_{n-1} * p_n)(k)$ for the distribution of S_n . An easy induction hence shows that $q_n(k) = (p_1 * \dots * p_n)(k)$ so that *the distribution of S_n is the n -fold convolution of the marginal distributions p_1, \dots, p_n .* As a corollary we observe that the convolution of any pair of distributions *must necessarily be a distribution*.

3 Towards the continuum

At scattered instances in our discussion of discrete probability experiments we have discovered that probabilities may be approximated by integrals of the form

$$\mathbf{P}\{a < X < b\} \approx \int_a^b f(x) dx$$

asymptotically in some parameter of the problem.

EXAMPLES: 1) *Limiting uniform experiments.* For each $n \geq 1$, suppose X_n taking values in the discrete set of points $\{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}\}$ with $\mathbf{P}\{X_n = k/n\} = 1/n$ for each $0 \leq k \leq n-1$. The points k/n pepper the unit interval increasingly densely as n increases and in the limit we pass to a continuum of values. The number of the points k/n encountered in any given subinterval (a, b) differs from $\frac{b}{1/n} - \frac{a}{1/n}$ only in the addition or deletion of a point at either boundary. These irritating round-off factors vanish on dividing by n and accordingly,

$\mathbf{P}\{a < X_n < b\} \rightarrow b - a$ as $n \rightarrow \infty$. If we set $u(x) := 1$ for x in the unit interval $(0, 1)$ then we may write $\mathbf{P}\{a < X_n < b\} \rightarrow \int_a^b u(x) dx$ where $u(x)$ stands in the rôle of a limiting probability density on the unit interval. We encountered this passage to the limit in Chapter V: the function u is the *uniform density* and serves as a model for the generation of a random point in the unit interval.

2) *Limiting geometric probabilities.* Consider a sequence of coin tosses where the success probability $p = p_n$ decreases to zero as n increases in such a way that $np_n \rightarrow \alpha$ for some fixed α . For each positive integer k , suppose X_n takes value k/n with probability $(1 - p_n)^k p_n$. Then X_n represents a suitably scaled waiting time till the first success in a succession of tosses, the natural scaling by n ensuring that the expected waiting time till the first success is bounded. As the approximation $(1 - x/n)^n \approx e^{-x}$ gets increasingly good as $n \rightarrow \infty$, in the limit of large n we obtain³

$$\mathbf{P}\{X_n \leq x\} = 1 - \sum_{k:k>nx} (1 - p_n)^k p_n = 1 - (1 - p_n)^{\lfloor nx+1 \rfloor} \rightarrow 1 - e^{-\alpha x},$$

and it follows that $\mathbf{P}\{a < X_n < b\} \rightarrow e^{-\alpha a} - e^{-\alpha b}$. If we write $w(x) := \alpha e^{-\alpha x}$ for $x \geq 0$, we hence obtain $\mathbf{P}\{a < X_n < b\} \rightarrow \int_a^b w(x) dx$ and, again, the function $w(x)$ may be seen in the light of a limiting probability density for a scaled geometric random variable with support now on the entire half-line. This is the *exponential density with parameter α* . As we shall see, the exponential density serves in many applications as a model of randomness because of its characteristic memoryless property.

3) *The de Moivre–Laplace limit law.* Let S_n be the number of successes in n tosses of a coin with (fixed) success probability p . Then S_n has the binomial distribution $\mathbf{P}\{S_n = k\} = \binom{n}{k} p^k (1 - p)^{n-k}$ with mean np and variance $np(1 - p)$. The shifted and scaled variable $S_n^* = (S_n - np)/\sqrt{np(1 - p)}$ takes values in the set $\{(k - np)/\sqrt{np(1 - p)}, 0 \leq k \leq n\}$ and as n increases these values pepper the real line increasingly finely. The de Moivre–Laplace theorem of Section VI.5 (or, more precisely, Problem VI.11) asserts that

$$\mathbf{P}\{a < S_n^* < b\} \rightarrow \int_a^b \phi(x) dx$$

where the function $\phi(x) := e^{-x^2/2}/\sqrt{2\pi}$ represents the limiting density of S_n^* on the real line. This is the important *standard normal density* ubiquitous in applications and the theory by virtue of the central limit theorem.

4) *The arc sine laws.* Let Z_{2n} denote the location of the last return to zero for a random walk over $2n$ unit (positive or negative) steps starting at the origin. We

³As is usual, $\lfloor x \rfloor$ denotes the greatest integer no larger than x .

shall see in Section VIII.5 that, for $0 < a < b < 1$,

$$\mathbf{P}\left\{a < \frac{1}{2^n} Z_{2n} < b\right\} \rightarrow \int_a^b s(x) dx$$

where $s(x) = \pi^{-1}x^{-1/2}(1-x)^{-1/2}$ represents the limiting density of $\frac{1}{2^n} Z_{2n}$ on the unit interval $0 < x < 1$. This is the *arc sine density* encountered in fluctuation theory. ▶

This kind of formal passage to the continuum through a simple discrete model has the virtue of being a useful exercise for beginners and helps solidify intuition. Eventually, however, it is simpler conceptually—and more intuitive—to deal directly with the continuous space.

4 Densities in one dimension

The continuous sample space $\mathbb{R} = (-\infty, \infty)$ was introduced in Section I.7. We begin with a summary of the salient features.

Probability masses give way naturally to densities in the continuum. Formally, a positive, integrable function $f: \mathbb{R} \rightarrow \mathbb{R}^+$ is a *density*⁴ if it is normalised so that $\int_{-\infty}^{\infty} f(x) dx = 1$. Events now correspond to subsets \mathbb{A} of the continuum. A chance variable X taking a continuum of real values is drawn according to f (or has density f) if, for each subset \mathbb{A} of the line, the probability that X takes values in \mathbb{A} is obtained as the integral of f over \mathbb{A} or, in notation,

$$\mathbf{P}\{X \in \mathbb{A}\} = \int_{\mathbb{A}} f(x) dx.$$

We will for the time being deal exclusively with piecewise continuous densities f and sets \mathbb{A} that are intervals or finite collections of intervals. In this case the usual interpretation of the integral in the sense of Riemann as an “area under the curve” causes no difficulty. Eventually, however, we are going to have to worry about what an expression like $\int_{\mathbb{A}} f(x) dx$ means for an arbitrary set \mathbb{A} and we will return to this question in Chapter XIII.

The *distribution function* of X (abbreviated *d.f.*) is the function

$$F(x) = \mathbf{P}\{X \leq x\} = \int_{-\infty}^x f(\xi) d\xi.$$

It is clear that $F(x)$ is a continuous, monotonically increasing function of x with limiting values of 0 as $x \rightarrow -\infty$ and 1 as $x \rightarrow +\infty$. There is no great harm in writing $F(-\infty) = 0$ and $F(+\infty) = 1$, where by $F(\pm\infty)$ we mean the limit of $F(x)$

⁴The elaboration *probability density function* is an unnecessary flourish in the context.

as $x \rightarrow \pm\infty$, and we do so without further ado. While the density determines the distribution function, the converse is also true and we can determine the density (at least at points of continuity) from the distribution function by differentiation, $f(x) = F'(x) = dF(x)/dx$, via the fundamental theorem of integral calculus.

In accordance with the usual terminology for functions, we say that a density $f(x)$ has *support* \mathbb{A} if $f(x) = 0$ for $x \notin \mathbb{A}$. (It is also permissible to say that f is *concentrated on* \mathbb{A} .) We will take the notational liberty in such cases of saying that the random variable X whose density is f has support in \mathbb{A} or that the distribution of X has support \mathbb{A} . Here and subsequently we adopt the convention that we only specify densities over their support in order to avoid the frequent caveats for completeness that would otherwise encumber the notation. *When densities are specified only in a range they are to be supposed to be identically zero outside it.* In this context, random variables whose density has support in the positive half-axis $0 \leq x < \infty$ play an important rôle. If X is a random variable of this type then it will, with probability one, take only positive values—we say then that X is a *positive random variable*—and it is only needful to specify its density $f(x)$ for $x \geq 0$.

In applications one not infrequently encounters a density, say \tilde{f} , that is related to a given density f by a shift of the origin and possibly a scaling of the axis, that is to say, $\tilde{f}(x) = \frac{1}{a} f\left(\frac{x-m}{a}\right)$ for some real value m and positive value a . (The change of variable $t = (x - m)/a$ allows an easy verification that \tilde{f} is indeed a density.) The corresponding distribution functions are then related via $\tilde{F}(x) = F\left(\frac{x-m}{a}\right)$. In this case we say that the densities \tilde{f} and f (or distributions \tilde{F} and F , or random variables \tilde{X} and X) belong to the same *type*. The quantity m is referred to as the *centring parameter* and corresponds to a shift of the origin of the graph of $f(x)$ to the point m while a is the *scale parameter* which results in a compression or dilation of the x -axis (depending on whether $a < 1$ or $a > 1$, respectively). In many applications only the type of the density matters and it will be convenient to suitably centre and scale it for analytical purposes.

The most important of the basic densities are the uniform density $u(x) = 1$ ($0 < x < 1$), the exponential density $g(x) = e^{-x}$ ($x > 0$), the normal density $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$, and the associated densities of these types obtained by shifting and scaling that the reader has already encountered in Section I.7.

While the density of a random variable tells us all that there is to be known, statistically speaking, about it, other measures (cruder to be sure) may suffice to characterise it for the purposes of a given application. Of such measures, the most important is the *mean* or the *expectation*. In analogy with the discrete case, the mean of a continuous random variable X with density $f(x)$ (or, simply, the mean of X or, alternatively, the mean of f) is defined by

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

provided the integral converges absolutely.⁵ (In the event the integral diverges to $+\infty$ or $-\infty$ it does no great harm to say the mean is $+\infty$ or $-\infty$, respectively.) Of course, this is just the continuous analogue of the mean of an arithmetic random variable and, as in the discrete case, the notion is familiar from elementary physics as the *centre of mass*. We may hence think of the mean (when it exists) as the centre of probability mass of a random variable.

The mean, by itself, is a relatively crude measure. For instance, the densities $f_1(x) = 1$ for $-1/2 < x < 1/2$ and $f_2(x) = 1/100$ for $-50 < x < 50$ both have the same mean of 0, yet the latter is “spread out” rather more. The *variance* captures very roughly the idea of how far from the mean the values taken by the random variable are likely to range. Formally, if the random variable X has density $f(x)$ and mean $\mu = \mathbb{E}(X)$, its variance (as for the mean, we sometimes say the variance of the density f to mean the same thing) is defined by

$$\text{Var}(X) := \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

when the integral exists. With this proviso, expanding out the square inside the integral, $(x - \mu)^2 = x^2 - 2\mu x + \mu^2$, and integrating each term separately yields

$$\text{Var}(X) = \mathbb{E}(X^2) - \mu^2 \text{ where } \mathbb{E}(X^2) := \int_{-\infty}^{\infty} x^2 f(x) dx,$$

an equivalent alternative form for the variance that frequently simplifies computations. In this context it is usual to reserve the symbol $\sigma^2 = \text{Var}(X)$ for the variance of X (with the usual proviso on existence); the positive square-root $\sigma = \sqrt{\text{Var}(X)}$ is referred to as the *standard deviation* of X . The variance is not defined if the mean does not exist; when the mean is finite the variance is always positive though the integral may diverge to $+\infty$. As with the mean, the expression for the variance is familiar from mechanics; the variance is the probabilistic equivalent of the notion of the *moment of inertia* of a physical object.

Changes in centring and scale result in natural modifications to the mean and variance. Suppose the mean and variance corresponding to a density $f(x)$ are μ and σ^2 , respectively. Then the corresponding mean and variance of any other density $\tilde{f}(x) = \frac{1}{a}f(\frac{x-m}{a})$ of the same type are $\tilde{\mu} = a\mu + m$ and $\tilde{\sigma}^2 = a^2\sigma^2$, respectively. The proofs of these assertions require little more than a change of variable inside the respective integrals.

The mean and variance play a central and vital rôle in the theory. A fuller development of these concepts and their importance will begin with Chapter XIII.

More generally, a random variable is a function of some coordinate variable. What kinds of functions should be allowed for consideration? Well

⁵The integral may, unfortunately, not exist; the *Cauchy* density $f(x) = 1/(\pi(1+x^2))$ provides a simple example of a density without expectation.

surely, at a minimum, we would like to be able to consistently assign probabilities to events determined by these functions. The functions encountered in this chapter will be unexceptionable in this regard and will occasion no technical complications. We defer the general framework to Chapter XII.

Thus, beginning with a random variable X of a given distribution, we may form new random variables Y, Z, \dots , and these should be viewed as functions of the coordinate variable X . Events such as $\{Y \leq y\}$, $\{Z > z\}$ and the like may now be referred back to the originating coordinate variable. Each new random variable Y, Z, \dots defined in this fashion may be thought of as defining a basic coordinate variable in its own right with a distribution that is inherited from the distribution of the original coordinate variable X .

EXAMPLES: 1) *Variables of the same type.* The simplest derived variables are obtained via a change of origin and scale. For example, for any positive a and real m we may define a new random variable $X_{m,a} = m + aX$. Of course, this corresponds simply to a new centring and scaling. The event $\{X_{m,a} \leq x\}$ corresponds to the set of sample points $\{X \leq (x-m)/a\}$ so that $X_{m,a}$ has distribution function

$$F_{m,a}(x) := P\{X_{m,a} \leq x\} = P\left\{X \leq \frac{x-m}{a}\right\} = F\left(\frac{x-m}{a}\right).$$

Differentiation shows that the associated density is related to the density of X via $f_{m,a}(x) = \frac{1}{a} f\left(\frac{x-m}{a}\right)$ so that $X_{m,a}$ and X have densities of the same type. As we have seen, the mean and variance of the derived variable $X_{m,a}$ may be related to that of X via $E(X_{m,a}) = m + a E(X)$ and $\text{Var}(X_{m,a}) = a^2 \text{Var}(X)$.

2) *Squared transformation.* If $Y = X^2$ then, for $y > 0$, the event $\{Y \leq y\}$ corresponds to the set of sample points $\{-\sqrt{y} \leq X \leq \sqrt{y}\}$ and it follows that the d.f., say, $G(y)$ of Y is given by

$$\begin{aligned} G(y) &= P\{Y \leq y\} = P\{-\sqrt{y} \leq X \leq \sqrt{y}\} = \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx \\ &= \int_{-\infty}^{\sqrt{y}} f(x) dx - \int_{-\infty}^{-\sqrt{y}} f(x) dx = F(\sqrt{y}) - F(-\sqrt{y}) \quad (y > 0). \end{aligned}$$

Of course, $G(y)$ is zero for $y \leq 0$. Formal differentiation shows now that Y is a positive random variable with density

$$g(y) = \frac{1}{2\sqrt{y}} [f(\sqrt{y}) + f(-\sqrt{y})] \quad (y > 0). \quad (4.1)$$

3) *Inverse transformation.* Suppose $T = 1/X$. If $t > 0$, the event $T \leq t$ occurs when $0 < X \leq 1/t$; and, if $t < 0$, the event $T \leq t$ occurs when $1/t \leq X < 0$. The associated d.f., say, $H(t)$ of T is hence given by

$$H(t) = P\{T \leq t\} = \begin{cases} 1 - \int_0^{1/t} f(x) dx & \text{if } t > 0, \\ \int_{1/t}^0 f(x) dx & \text{if } t < 0. \end{cases}$$

Formal differentiation shows that the associated density of T is given by $h(t) = t^{-2}f(t^{-1})$ for $t \neq 0$. Probability calculations are unaffected by the addition or deletion of individual points at the boundary of a Riemann integral and so the value of $h(t)$ at $t = 0$ may be set to any convenient value. ▶

It is not difficult to craft a quite general rule to determine the densities of transformed variables when the transformation is monotonically increasing or decreasing. Suppose the derived variable Y is related to the coordinate variable X via the relation $X = \zeta(Y)$ where the map ζ is either continuously increasing or continuously decreasing over the support of X . Let $f(x)$ be the density of X and $g(y)$ the density of Y . Then $g(y) = f(\zeta(y))|\zeta'(y)|$.

The proof is little more than an exercise in change of variable of integration. To any set of points \mathbb{A} let $\zeta^{-1}\mathbb{A}$ denote the preimage under ζ of \mathbb{A} , that is to say, $\zeta^{-1}\mathbb{A}$ is the set of points $\{y : \zeta(y) \in \mathbb{A}\}$. The usual change of variable formula for integration then shows that the event $\{X \in \mathbb{A}\}$ has probability

$$\mathbf{P}\{X \in \mathbb{A}\} = \int_{\mathbb{A}} f(x) dx = \int_{\zeta^{-1}\mathbb{A}} f(\zeta(y)) |\zeta'(y)| dy.$$

On the other hand,

$$\mathbf{P}\{X \in \mathbb{A}\} = \mathbf{P}\{Y \in \zeta^{-1}\mathbb{A}\} = \int_{\zeta^{-1}\mathbb{A}} g(y) dy.$$

As these relations hold for any choice of event \mathbb{A} , a comparison of the right-hand sides of the two relations proves the desired relation between the densities.

EXAMPLES: 4) If $Y = e^X$ (or, equivalently, $X = \log Y$) then Y is a positive variable with density $g(y) = f(\log y)/y$; more generally, if $Y = a^X$ for some positive a then $g(y) = \frac{1}{y|\log a|} f\left(\frac{\log y}{\log a}\right)$ for $y > 0$.

5) If X is positive and $Y = \log X$ then $g(y) = e^y f(e^y)$.

6) If X has support in the interval $[0, \pi]$ then $Y = \cos X$ has support in the interval $[-1, 1]$ and density $g(y) = f(\arccos y)/\sqrt{1-y^2}$.

7) If X is positive then $Y = \sqrt{X}$ is also positive with density $2yf(y^2)$; more generally, if $r > 0$ then $Y = X^r$ has density $g(y) = \frac{1}{r}y^{-1+1/r}f(y^{1/r})$. ▶

5 Densities in two and more dimensions

The Euclidean plane serves as the model sample space when one may naturally identify a pair of random variables as the outcome of a conceptual experiment. A triple of random variables may be identified in like fashion as specifying a sample point in three-dimensional space and, more generally, the sample space

for an abstract experiment that results in an n -tuple of random variables may be taken to be Euclidean space \mathbb{R}^n . The mathematical considerations in such settings are not materially different from those encountered in one dimension—though the notation is inescapably more cumbrous. One new feature that does emerge is in the possible interactions between the variables.

It is simplest to deal with two dimensions; matters don't change materially, more complex notation excepting, in three or more dimensions.

Densities in two dimensions naturally acquire units of probability mass per unit area. Formally, a density on the Euclidean plane is a positive, integrable function $f: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+$ normalised so that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1.$$

A pair of continuous-valued random variables (X_1, X_2) has density f (or is drawn according to f) if, for any sufficiently regular region \mathbb{A} in the plane, the probability that the random pair (X_1, X_2) lies in \mathbb{A} is given by

$$P\{(X_1, X_2) \in \mathbb{A}\} = \iint_{\mathbb{A}} f(x_1, x_2) dx_1 dx_2.$$

Again, for the time being, we will deal exclusively with regions \mathbb{A} that are rectangles, discs, or other geometrically regular shapes, and finite unions of such sets. In such cases it is unambiguous what the integral means; thus, for instance, the event that (X_1, X_2) takes values in the rectangle $(a_1, b_1] \times (a_2, b_2]$ is equivalent to the statement that the inequalities $a_1 < X_1 \leq b_1$ and $a_2 < X_2 \leq b_2$ simultaneously hold and the probability of this event is then given by

$$P\{a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2\} = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_2 dx_1.$$

Of course, by additivity, the probability of a disjoint union of such events is equal to the sum of expressions of the above type.

The (two-dimensional) d.f. associated with the density f is the function

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(\xi, \eta) d\eta d\xi.$$

As for the case in one dimension, we may recover the density (wherever continuous) from the d.f. by differentiation,

$$f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2},$$

so that either the density or the associated d.f. suffices to specify the distribution of the random pair (X_1, X_2) .

The event $\{X_1 \leq x_1\}$ refers to the region in the plane defined by the collection of points (ξ, η) satisfying $-\infty < \xi \leq x_1$ and $-\infty < \eta < \infty$. The distribution of the coordinate variable X_1 alone is hence specified by the function

$$F_1(x_1) = P\{X_1 \leq x_1\} = F(x_1, +\infty) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f(\xi, \eta) d\eta d\xi.$$

Differentiation then shows that the density of X_1 is given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, \eta) d\eta.$$

The integral on the right conjures up the mental picture of f_1 as obtained by collapsing the density $f(\xi, \eta)$ along the vertical line $\xi = x_1$ by integrating out the second variable to the “margin” of the page; in view of this imagery the one-dimensional density f_1 and the associated d.f. F_1 are called *marginals*. We may now compute the mean and the variance of X_1 (if they exist) by the usual formulations

$$\begin{aligned} \mu_1 &= E(X_1) = \int_{-\infty}^{\infty} x_1 f_1(x_1) dx_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1, x_2) dx_2 dx_1, \\ \sigma_1^2 &= \text{Var}(X_1) = \int_{-\infty}^{\infty} (x_1 - \mu_1)^2 f_1(x_1) dx_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1)^2 f(x_1, x_2) dx_2 dx_1. \end{aligned}$$

Entirely similar considerations show that the marginal d.f. of the coordinate variable X_2 is given by

$$F_2(x_2) = P\{X_2 \leq x_2\} = F(+\infty, x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{\infty} f(\xi, \eta) d\xi d\eta$$

(it being natural now to swap the order of integration) with the associated marginal density

$$f_2(x_2) = \int_{-\infty}^{\infty} f(\xi, x_2) d\xi$$

obtained by integrating out the first variable. The mean μ_2 and variance σ_2^2 of X_2 are obtained by the usual integrations of the marginal density f_2 .

The special case when f may be separated into a product of its marginals f_1 and f_2 is of particular importance. We say that the associated random variables X_1 and X_2 are *independent* if $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. How does this notion of independence square with the earlier notion of independent events?

Suppose A_1 and A_2 are finite unions of intervals on the real line. The Cartesian product of A_1 and A_2 , denoted $A_1 \times A_2$, is called a *rectangle* in the plane \mathbb{R}^2 . Figure 1 illustrates the idea. The sets of points A_1 and A_2 , respectively, correspond to the (union of the) intervals indicated in bold along each of the axes; the rectangle $A_1 \times A_2$ is the shaded region in the plane. Of course,

when \mathbb{A}_1 and \mathbb{A}_2 are intervals then $\mathbb{A}_1 \times \mathbb{A}_2$ forms an ordinary rectangle. The event $\{X_1 \in \mathbb{A}_1, X_2 \in \mathbb{A}_2\}$ corresponds to the set of sample points (X_1, X_2) lying in the rectangle $\mathbb{A}_1 \times \mathbb{A}_2$ and consequently

$$\begin{aligned} \mathbf{P}\{X_1 \in \mathbb{A}_1, X_2 \in \mathbb{A}_2\} &= \iint_{\mathbb{A}_1 \times \mathbb{A}_2} f(x_1, x_2) dx_2 dx_1 = \int_{\mathbb{A}_1} \int_{\mathbb{A}_2} f_1(x_1) f_2(x_2) dx_2 dx_1 \\ &= \left(\int_{\mathbb{A}_1} f_1(x_1) dx_1 \right) \left(\int_{\mathbb{A}_2} f_2(x_2) dx_2 \right) = \mathbf{P}\{X_1 \in \mathbb{A}_1\} \mathbf{P}\{X_2 \in \mathbb{A}_2\}. \end{aligned}$$

Thus, if X_1 and X_2 are independent then so are the events $\{X_1 \in \mathbb{A}_1\}$ and $\{X_2 \in \mathbb{A}_2\}$ for any choices of \mathbb{A}_1 and \mathbb{A}_2 .

Independence places a peculiar and powerful constraint on the underlying functions as the one-dimensional marginals f_1 and f_2 completely determine the two-dimensional density f . It is clear that this is not true in general. In a general setting, what can be said about the level of dependence between two variables? Suppose X_1 and X_2 have means μ_1 and μ_2 , respectively, and variances σ_1^2 and σ_2^2 , respectively. In analogy with the concept of the variance, we may define the *covariance* between X_1 and X_2 by

$$\text{Cov}(X_1, X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2) dx_2 dx_1. \quad (5.1)$$

Expanding the product inside the integrand and evaluating the resulting integrals one at a time we may write the covariance in the form $\text{Cov}(X_1, X_2) = E(X_1 X_2) - \mu_1 \mu_2$, where

$$E(X_1 X_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f(x_1, x_2) dx_2 dx_1 \quad (5.2)$$

is called the *correlation* between X_1 and X_2 and connotes the expected value of the product. In the special case when X_1 and X_2 are independent the integrals may be evaluated separately and the right-hand side of (5.1) evaluates to

$$\int_{-\infty}^{\infty} (x_1 - \mu_1) f_1(x_1) dx_1 \int_{-\infty}^{\infty} (x_2 - \mu_2) f_2(x_2) dx_2 = 0.$$

We say that the random variables X_1 and X_2 are *uncorrelated* if $\text{Cov}(X_1, X_2) = 0$ or, equivalently, $E(X_1 X_2) = E(X_1) E(X_2)$. Thus, *independent random variables are uncorrelated*. Consequently, if the covariance is non-zero then the variables are dependent and the covariance provides a simple numerical test for dependence. The converse is not true, in general, as we will see in the examples to follow: *uncorrelated variables are not necessarily independent*.

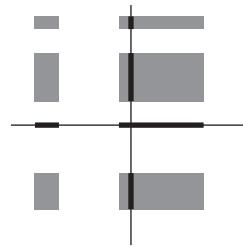


Figure 1: Rectangle.

EXAMPLES: 1) *The uniform distribution in the unit square.* The density $f(x_1, x_2)$ which takes value 1 inside the unit square $\{(x_1, x_2) : 0 < x_1 < 1, 0 < x_2 < 1\}$ and value 0 outside it represents a chance experiment producing a random point in the unit square. The associated marginal densities are easily seen to be $f_1(x_1) = u(x_1)$ and $f_2(x_2) = u(x_2)$ where $u(x) = 1$ for $0 < x < 1$. Accordingly, $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ and the variables X_1 and X_2 are independent, each distributed uniformly in the unit interval.

2) *The uniform distribution in the unit disc.* The density f with support in the unit disc and defined by $f(x_1, x_2) = \pi^{-1}$ for $x_1^2 + x_2^2 < 1$ represents a chance experiment producing a random point in the unit disc. When $|x_1| < 1$, along the line $\xi = x_1$ the function $f(\xi, \eta)$ is equal to $1/\pi$ only for $|\eta| < (1 - x_1^2)^{1/2}$ (and is zero elsewhere). The marginal density of X_1 is hence easily seen to be given by $f_1(x_1) = (2/\pi)(1 - x_1^2)^{1/2}$ for $-1 < x_1 < 1$. By symmetry, the marginal density f_2 is exactly of the same form. An easy computation now shows that X_1 and X_2 both have zero mean and variance $1/4$ (a trigonometric substitution renders the latter computation transparent). The bilinear form $\xi\eta$ alternates in sign over each of the four quadrants whence $\text{Cov}(X_1, X_2) = \frac{1}{\pi} \iint_{\xi^2 + \eta^2 < 1} \xi\eta \, d\eta d\xi = 0$ by skew symmetry and no integrations are required. It follows that X_1 and X_2 are uncorrelated but not independent as $f(0, 0) = 1/\pi \neq 4/\pi^2 = f_1(0)f_2(0)$.

3) *The uniform distribution on the unit sphere.* Cartographers are familiar with the representation of the points on the unit sphere in three dimensions in terms of geographic longitude φ and latitude ϑ via the equations $x = \cos(\vartheta)\cos(\varphi)$, $y = \cos(\vartheta)\sin(\varphi)$, and $z = \sin(\vartheta)$ with $-\pi/2 < \vartheta < \pi/2$ and $-\pi < \varphi \leq \pi$. (The description omits the two poles but isolated points will end up with probability zero in our model in any case and the omission of the poles need not worry us.) As the unit sphere in three dimensions has surface area 4π , the conceptual experiment of selecting a random point on the surface of the sphere should attach to any region \mathbb{A} on the sphere the probability $\text{Area}(\mathbb{A})/4\pi$. Now, at any latitude ϑ , the area covered by an infinitesimal strip traversing a differential latitude $d\vartheta$ and longitude $d\varphi$ is given by $\cos(\vartheta)d\varphi d\vartheta$. The selection of a random point on the surface is hence equivalent to specifying a latitude and longitude in accordance with the density $f(\vartheta, \varphi) = (4\pi)^{-1}\cos(\vartheta)$ for $-\pi/2 < \vartheta < \pi/2$ and $-\pi < \varphi \leq \pi$. The corresponding marginal densities are $f_1(\vartheta) = \cos(\vartheta)/2$ and $f_2(\varphi) = 1/2\pi$ (over the appropriate ranges for ϑ and φ) and the coordinate latitude and longitude are independent variables.

4) *The Cauchy density.* The function defined by $f(x_1, x_2) = \frac{1}{2\pi}(1 + x_1^2 + x_2^2)^{-3/2}$ is the Cauchy density on the plane. Integrating out one variable at a time [the change of variable $x_2 = (1 + x_1^2)^{1/2}\tan(\theta)$ is perspicuous] shows that the marginal densities have the same form and are given by $f_1(x_1) = 1/(\pi(1 + x_1^2))$ and $f_2(x_2) = 1/(\pi(1 + x_2^2))$; these are the standard Cauchy densities on the line. It follows that the corresponding marginal variables are not independent.

5) *The bivariate normal density.* Suppose $-1 < \rho < 1$. In a slight abuse of notation we introduce the positive function

$$\phi(x_1, x_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-(x_1^2 - 2\rho x_1 x_2 + x_2^2)}{2(1-\rho^2)}\right), \quad (5.3)$$

the quadratic form in the exponent reminiscent of the one-dimensional normal density $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$. That it is a density may be verified by completing squares in the exponent, $x_1^2 - 2\rho x_1 x_2 + x_2^2 = (1-\rho^2)x_1^2 + (x_2 - \rho x_1)^2$, and factoring the exponential form into the product

$$\phi(x_1, x_2; \rho) = \phi(x_1) \cdot \frac{1}{\sqrt{1-\rho^2}} \phi\left(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}}\right).$$

Integrating out the variables sequentially yields

$$\begin{aligned} \iint_{-\infty}^{\infty} \phi(x_1, x_2; \rho) dx_2 dx_1 &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \frac{1}{\sqrt{1-\rho^2}} \phi\left(\frac{x_2 - \rho x_1}{\sqrt{1-\rho^2}}\right) dx_2 \right\} \phi(x_1) dx_1 \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \phi(u) du \right\} \phi(x_1) dx_1, \end{aligned}$$

the change of variable $u = (x_2 - \rho x_1)/\sqrt{1-\rho^2}$ in the inner integral reducing the right-hand side to a product of two integrals of the standard normal density $\phi(\xi) = (2\pi)^{-1/2} e^{-\xi^2/2}$. As, by Lemma VI.1.1, the standard normal has unit area, it follows that, indeed, $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x_1, x_2; \rho) dx_2 dx_1 = 1$. If σ_1 and σ_2 are strictly positive and μ_1 and μ_2 real, a shift and scale of the coordinate axes leads to a density $\frac{1}{\sigma_1 \sigma_2} \phi\left(\frac{x_1 - \mu_1}{\sigma_1}, \frac{x_2 - \mu_2}{\sigma_2}; \rho\right)$ of the same type.

As a by-product of the calculation, by integrating out only the second variable, we obtain the marginal density $f_1(x_1) = \phi(x_1)$ and, by an entirely analogous calculation, $f_2(x_2) = \phi(x_2)$. Thus, *the densities in the (bivariate) normal family $\phi(\cdot, \cdot; \rho)$ with $-1 < \rho < 1$ all have (univariate) normal marginals.* The reader should resist the lure of the tempting converse: *random variables that have marginal normal densities are not necessarily (jointly) normal.* See Problem 2.

The same change of variable $x_2 = \rho x_1 + u\sqrt{1-\rho^2}$ shows that

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 \phi(x_1, x_2; \rho) dx_2 dx_1 &= \int_{-\infty}^{\infty} x_1 \phi(x_1) \\ &\times \left\{ \int_{-\infty}^{\infty} (\rho x_1 + u\sqrt{1-\rho^2}) \phi(u) du \right\} dx_1 = \rho \int_{-\infty}^{\infty} x_1^2 \phi(x_1) dx_1 = \rho \end{aligned}$$

by two uses of Lemma VI.1.2. It follows that $\text{Cov}(X_1, X_2) = \rho$. As $\phi(x_1, x_2; \rho) = \phi(x_1)\phi(x_2)$ if, and only if, $\rho = 0$, *a necessary and sufficient condition for a bivariate normal pair to be independent is that they be uncorrelated.* ►

The expression (5.2) may be interpreted as the expectation of the product $X_1 X_2$. More generally, the formal expression

$$E(g(X_1, X_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) f(x_1, x_2) dx_1 dx_2$$

denotes the expectation of the variable $g(X_1, X_2)$ provided g is sufficiently regular that the expectation integral exists. In particular, by selecting $g(X_1, X_2) = X_1 + X_2$ an easy computation using the linearity of integration shows that *expectation is additive*: $E(X_1 + X_2) = E(X_1) + E(X_2)$. If, additionally, X_1 and X_2 are independent then, by selecting $g(X_1, X_2) = ((X_1 - \mu_1) + (X_2 - \mu_2))^2$ and expanding the indicated square, linearity of integration shows again that *variance is additive if the summands are independent*: $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$. I will leave the easy demonstrations to the reader.

The generalisation of these ideas to more than two coordinate variables is straightforward. A positive and integrable function $f: \mathbb{R}^n \rightarrow \mathbb{R}^+$ is a density (in n dimensions) if $\int \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) dx_n \dots dx_1 = 1$. We say that a sequence of n random variables (X_1, \dots, X_n) is distributed according to f if, for all suitably regular regions A in \mathbb{R}^n , the probability that (X_1, \dots, X_n) takes values in A is given by

$$P\{(X_1, \dots, X_n) \in A\} = \int \cdots \int_A f(x_1, \dots, x_n) dx_n \dots dx_1. \quad (5.4)$$

The distribution function associated with f is defined by

$$F(x_1, \dots, x_n) = P\{X_1 \leq x_1, \dots, X_n \leq x_n\} = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(\xi_1, \dots, \xi_n) d\xi_n \cdots d\xi_1,$$

and, as before, the density may be recovered from its d.f. by successive partial derivatives.

As is the case for two variables, for each j , the marginal density $f_j(x_j)$ of the variable X_j may be obtained by integrating out all the other variables in the density f .

Convention. Vector notation helps simplify presentation when there are more than two dimensions. We use lowercase bold font letters to denote vector-valued variables, $\mathbf{x} = (x_1, \dots, x_n)$ denoting a generic point in \mathbb{R}^n , and uppercase bold font letters to denote random vectors, $\mathbf{X} = (X_1, \dots, X_n)$ denoting a generic random vector in \mathbb{R}^n . For definiteness, we interpret \mathbf{x} and \mathbf{X} as *row vectors* in matrix–vector operations (this particular stylistic choice inspired purely by my desire to keep equations compact); the transposes \mathbf{x}^T and \mathbf{X}^T represent the associated *column vectors*. Vector inequalities such as $\mathbf{x} \leq \mathbf{y}$ are to be interpreted componentwise, $x_j \leq y_j$ for each j .

In vector notation (5.4) takes the economical form

$$P\{\mathbf{X} \in A\} = \int_A f(\mathbf{x}) d\mathbf{x}, \text{ and, in particular, } F(\mathbf{x}) = \int_{\xi \leq \mathbf{x}} f(\boldsymbol{\xi}) d\boldsymbol{\xi}. \quad (5.4')$$

The n means of the component variables may now be written compactly in terms of the n -dimensional *mean vector* $\mathbf{E}(\mathbf{X}) = \boldsymbol{\mu} = [\mu_j]$ where $\mu_j = \mathbf{E}(X_j)$ for $1 \leq j \leq n$. Likewise, the system of n coordinate variables has associated with it n variances and $\binom{n}{2}$ covariances which are best represented as an $n \times n$ *covariance matrix* $\text{Cov}(\mathbf{X}) = \mathbf{A} = [A_{jk}]$ whose diagonal elements are the variances, $A_{jj} = \text{Var}(X_j)$, and the off-diagonal elements are the covariances, $A_{jk} = \text{Cov}(X_j, X_k)$ if $j \neq k$. Interpreting expectations of matrices componentwise we may write the covariance matrix in terms of matrix products via $\text{Cov}(\mathbf{X}) = \mathbf{E}((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T) = \mathbf{E}(\mathbf{X}^T \mathbf{X}) - \mathbf{E}(\mathbf{X}^T) \mathbf{E}(\mathbf{X})$.

The random variables X_1, \dots, X_n are independent if, and only if, their (n -dimensional) density may be factored into a product of their marginal densities, that is to say, $f(x_1, \dots, x_n) = f_1(x_1) \cdots f_n(x_n)$. In particular, suppose \mathbb{A}_j is a finite union of intervals for each j . The Cartesian product $\mathbb{A} = \mathbb{A}_1 \times \cdots \times \mathbb{A}_n$ is called an (n -dimensional) *rectangle* and the event $\{\mathbf{X} \in \mathbb{A}\}$ occurs if, and only if, $X_j \in \mathbb{A}_j$ for each j . It follows just as in the case of two dimensions that if X_1, \dots, X_n are independent, then $P\{\mathbf{X} \in \mathbb{A}\} = \prod_{j=1}^n P\{X_j \in \mathbb{A}_j\}$. Thus *independence of random variables is equivalent to a rule of product of probabilities over rectangles*.

A key observation may be made at this point. Suppose $X_1, \dots, X_n, Y_1, \dots, Y_m$ are independent random variables grouped into two sets of n and m variables apiece for notational convenience. Suppose U and V are random variables where U is a function solely of X_1, \dots, X_n and V is a function solely of Y_1, \dots, Y_m . Then U and V are independent. To see this, fix any u and v . Let $\mathbb{A}(u)$ be the set of n -tuples (x_1, \dots, x_n) for which $U(x_1, \dots, x_n) \leq u$ and let $\mathbb{B}(v)$ be the set of m -tuples (y_1, \dots, y_m) for which $V(y_1, \dots, y_m) \leq v$. Then

$$\begin{aligned} P\{U \leq u, V \leq v\} &= P\{(X_1, \dots, X_n) \in \mathbb{A}(u), (Y_1, \dots, Y_m) \in \mathbb{B}(v)\} \\ &= P\{(X_1, \dots, X_n) \in \mathbb{A}(u)\} P\{(Y_1, \dots, Y_m) \in \mathbb{B}(v)\} = P\{U \leq u\} P\{V \leq v\} \end{aligned}$$

and it follows that U and V are independent. Extensions to more than two random variables U and V are straightforward: if X_1, \dots, X_n are independent random variables and U_1, \dots, U_k is any collection of k random variables each of which is determined by a non-overlapping subgroup of the X_i then U_1, \dots, U_k are independent.

6 Randomisation, regression

Two and more coordinate variables create the possibility of interaction that is abeyant when dealing with a single coordinate variable. Suppose the pair of coordinate variables (X_1, X_2) has a continuous density $f(x_1, x_2)$ and let $f_1(x_1)$ be the marginal density of X_1 . Suppose $f_1(x_1)$ is continuous and strictly positive in an interval containing a given point $x_1 = \xi$. Corresponding to each such ξ we may then define the positive function

$$v_\xi(\eta) = \frac{f(\xi, \eta)}{f_1(\xi)}. \quad (6.1)$$

Integration shows that $\int_{-\infty}^{\infty} v_{\xi}(\eta) d\eta = \frac{1}{f_1(\xi)} \int_{-\infty}^{\infty} f(\xi, \eta) d\eta = 1$, and it follows that $v_{\xi}(\cdot)$ is a density for each such ξ , the corresponding d.f. given by

$$V_{\xi}(\eta) = \int_{-\infty}^{\eta} v_{\xi}(x_2) dx_2 = \frac{1}{f_1(\xi)} \int_{-\infty}^{\eta} f(\xi, x_2) dx_2. \quad (6.2)$$

How should we interpret this distribution?

By conditioning on the event that X_1 takes values in a small neighbourhood of the point ξ we obtain

$$\mathbf{P}\{X_2 \leq \eta \mid \xi < X_1 \leq \xi + h\} = \frac{\int_{\xi}^{\xi+h} \int_{-\infty}^{\eta} f(x_1, x_2) dx_2 dx_1}{\int_{\xi}^{\xi+h} f_1(x_1) dx_1}.$$

If f is continuous in a neighbourhood containing ξ and h is small, the right-hand side is given approximately by

$$\frac{h \int_{-\infty}^{\eta} f(\xi, x_2) dx_2}{h f_1(\xi)},$$

the error in approximation of the order of h itself. By allowing h to tend to zero it follows hence that $\mathbf{P}\{X_2 \leq \eta \mid \xi < X_1 \leq \xi + h\} \rightarrow V_{\xi}(\eta)$. Viewed in this light it is natural to interpret $V_{\xi}(\eta)$ as the d.f. of X_2 conditioned upon the event that X_1 takes values in an infinitesimally small neighbourhood of the point ξ ; we say that $V_{\xi}(\eta)$ is the *conditional distribution of X_2 given that $X_1 = \xi$* .

The reader may feel a little queasy about the fact that one is apparently conditioning on an event $X_1 = \xi$ of zero probability and worry about the constraints on the density f that are needed to pull off this kind of limiting sleight of hand. She will come to no harm if she bears firmly in mind that the distribution at hand is described precisely enough via the formulations (6.1,6.2), the limiting arguments merely attached *post hoc* to provide some intuition for these objects.

EXAMPLE 1) A paradox. To illustrate the dangers inherent in conditioning on sets of zero probability, consider a random pair (X_1, X_2) selected by uniform sampling from the unit disc in the plane. What can be said about the distribution of X_2 conditioned on the event $X_1 = X_2$? We can consider two limiting arguments.

The event A_h that $|X_1 - X_2| \leq h$ consists of the points in the unit disc in a diagonal band of cross-section $2h$ as shown in Figure 2(a). For $|\eta| < 1/\sqrt{2}$, the conditional probability that $X_2 \leq \eta$ given the occurrence of A_h is given by the ratio of the shaded area of the strip to the total area of the strip. Elementary geometry shows then that

$$\mathbf{P}\{X_2 \leq \eta \mid A_h\} = \frac{(2h + 2h \cdot \sqrt{2}\eta)/\pi}{4h/\pi} + o(h)$$

where the term $o(h)$ represents a quantity that tends to zero as $h \rightarrow 0$. By letting h tend to zero on the right-hand side we obtain the d.f. $V(\eta) = \frac{1}{2}(1 + \sqrt{2}\eta)$, whence taking

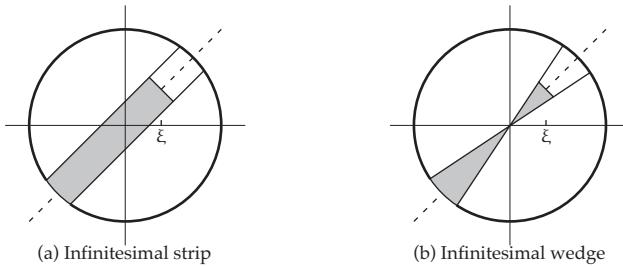


Figure 2: (a) An infinitesimal strip of width $2h$ around the line $x_1 = x_2$. (b) An infinitesimal wedge of arc length $2h$ around the line $x_1 = x_2$.

derivatives yields the uniform density $v(\eta) = 1/\sqrt{2}$ (for $|\eta| < 1/\sqrt{2}$) as a candidate for the conditional density of X_2 given that $X_1 = X_2$.

On the other hand, if we condition on the event B_h that (X_1, X_2) lies in the wedge-shaped regions with arc lengths $2h$ on either side, then elementary geometrical considerations show that

$$P\{X_2 \leq \eta \mid B_h\} = \begin{cases} \frac{h - \frac{1}{2} \cdot 2\eta^2 \cdot 2h}{2h} + o(h) & \text{if } -1/\sqrt{2} < \eta < 0, \\ \frac{h + \frac{1}{2} \cdot 2\eta^2 \cdot 2h}{2h} + o(h) & \text{if } 0 \leq \eta < 1/\sqrt{2}. \end{cases}$$

By taking the limit as $h \rightarrow 0$ of both sides, we obtain the d.f.

$$W(\eta) = \begin{cases} \frac{1}{2}(1 - 2\eta^2) & \text{if } -1/\sqrt{2} < \eta < 0, \\ \frac{1}{2}(1 + 2\eta^2) & \text{if } 0 \leq \eta < 1/\sqrt{2}, \end{cases}$$

and, taking derivatives, we obtain the function $w(\eta) = 2|\eta|$ (for $|\eta| < 1/\sqrt{2}$) as our second candidate for the conditional density of X_2 given that $X_1 = X_2$.

Clearly, both expressions cannot be right. Our paradox is resolved by the understanding that the limiting probability depends upon the nature of the limiting argument: different approaches to a limit can conceivably yield rather different answers. In this case there is no single, unambiguous, approach to the zero probability set $X_1 = X_2$. The definitions (6.1,6.2) avoid the difficulties inherent in specifying the nature of the approach to the limit and the regularity conditions that must hold perforce. ▶

The interpretation of $v_\xi(\eta)$ as a *conditional density* suggests the more graphic (if more cumbersome) notation $f_2(\eta \mid X_1 = \xi) = v_\xi(\eta)$. Viewed in this light, the connection with the definition of conditional probability is clear with probabilities replaced by densities in the defining ratio (6.1) when one segues to the continuum. The reader should beware, however, that, while vivid, the new notation does give rise to the wholly spurious impression that the marginal density $f_2(\cdot)$ and the conditional density $f_2(\cdot \mid X_1 = \xi)$ are somehow intimately related; she should bear in mind that the function $f_2(\cdot \mid X_1 = \xi)$ is *defined* by the ratio on the right in (6.1) and has no *a priori* relation with $f_2(\cdot)$.

In the new notation the associated conditional d.f. is written naturally as $F_2(\eta | X_1 = \xi) = V_\xi(\eta)$.

The mean of the conditional density $f_2(\cdot | X_1 = \xi)$ (if it exists) is denoted $E(X_2 | X_1 = \xi)$ and given by the usual formulation

$$E(X_2 | X_1 = \xi) = \int_{-\infty}^{\infty} \eta f_2(\eta | X_1 = \xi) d\eta = \frac{1}{f_1(\xi)} \int_{-\infty}^{\infty} \eta f(\xi, \eta) d\eta.$$

The conditional variance $\text{Var}(X_2 | X_1 = \xi)$ (if it exists) is defined likewise as the variance of the density $f_2(\cdot | X_1 = \xi)$.

It will be convenient as a matter of convention to set $E(X_2 | X_1 = \xi) = 0$ at any point ξ at which $f_1(\xi) = 0$. As X_1 varies over all possibilities ξ , the values $E(X_2 | X_1 = \xi)$ then range over the reals, or, in other words, the range of values spanned by the conditional expectation determines a *function* on the coordinate sample space. This function is denoted $E(X_2 | X_1)$ and called the *regression of X_2 on X_1* . (The unfortunate terminology is due to Francis Galton and is, regrettably, immovable by reasons of tradition.) We interpret $E(X_2 | X_1)$ as a random variable which takes the value $E(X_2 | X_1 = \xi)$ at all sample points of the chance experiment on which the coordinate variable X_1 takes value ξ . The conditional variance $\text{Var}(X_2 | X_1)$ is defined similarly.

All these considerations carry over naturally if we reverse the direction of conditioning. Suppose $f_2(x_2)$ is continuous and strictly positive in an interval containing the point $x_2 = \eta$. Then the conditional density of X_1 given that $X_2 = \eta$ is defined by $f_1(\xi | X_2 = \eta) = f(\xi, \eta)/f_2(\eta)$, with $F_1(\xi | X_2 = \eta)$ the corresponding conditional d.f. The regression function $E(X_1 | X_2)$ and conditional variance $\text{Var}(X_1 | X_2)$ are defined by similar considerations.

EXAMPLES: 2) *Return to the uniform density on the unit disc.* In Example 5.2, if $|\xi| < 1$, we have $f_2(\eta | X_1 = \xi) = 1/(2\sqrt{1 - \xi^2})$ with support in $|\eta| < \sqrt{1 - \xi^2}$. Thus, given $X_1 = \xi$, the random variable X_2 is uniformly distributed in the interval $(-\sqrt{1 - \xi^2}, \sqrt{1 - \xi^2})$. As this density has zero mean and variance $(1 - \xi^2)/3$, the regression function is given by $E(X_2 | X_1) = 0$ and the conditional variance by $\text{Var}(X_2 | X_1) = \frac{1}{3}(1 - \xi^2)$.

3) *Return to the bivariate normal.* If (X_1, X_2) have the bivariate normal density $\phi(x_1, x_2; \rho)$ of Example 5.5 then, for each real ξ , the conditional density of X_2 given that $X_1 = \xi$ is $f_2(\eta | X_1 = \xi) = (1 - \rho^2)^{-1/2} \phi((1 - \rho^2)^{-1/2}(\eta - \rho\xi))$. This is the normal density with mean $\rho\xi$ and variance $1 - \rho^2 < 1$. The regression of X_2 on X_1 is hence given by $E(X_2 | X_1) = \rho X_1$. Thus, conditioned on X_1 , the value of X_2 is proportional to X_1 and has a smaller variance. ▶

Another interpretation of the nature of the conditioning is provided by a consideration of the marginal distributions. The marginal density of X_2 may be recast in the form

$$f_2(\eta) = \int_{-\infty}^{\infty} f(\xi, \eta) d\xi = \int_{-\infty}^{\infty} f_2(\eta | X_1 = \xi) f_1(\xi) d\xi. \quad (6.3)$$

Taking expectation with respect to the formulations of the marginal density on either side, by a natural interchange in the order of integration we obtain

$$\int_{-\infty}^{\infty} \eta \left\{ \int_{-\infty}^{\infty} f_2(\eta | X_1 = \xi) f_1(\xi) d\xi \right\} d\eta = \int_{-\infty}^{\infty} E(X_2 | X_1 = \xi) f_1(\xi) d\xi$$

which we may write in the compact and memorable form

$$E(X_2) = E(E(X_2 | X_1)). \quad (6.4)$$

The expressions on the right of (6.3,6.4) may be viewed as the continuous analogue of the theorem of total probability. If we view X_1 as a random parameter influencing the distribution of X_2 then we may interpret (6.3) as saying that the distribution of X_2 is obtained by averaging the parametric density $f_2(\eta | X_1 = \xi) = v_\xi(\eta)$ with respect to the random parameter ξ ; in other words, the marginal distribution of X_2 is obtained by *randomisation over the parameter* X_1 . The difference in emphasis is one of perspective depending on whether it is more natural in an application to specify the joint density $f(\xi, \eta)$ or to specify the conditional density $f_2(\eta | X_1 = \xi)$ together with the marginal density $f_1(\xi)$. We have seen both points of view represented in the applications of conditional probability in Chapter II.

These concepts have natural counterparts when the variables are arithmetic but in this case there is no difficulty in interpretation as the conditional distributions are ordinary conditional probabilities.

The naïve notion of conditional expectation developed in this section suffices for our needs in this volume. Extending these ideas to a general theory of conditional expectation needs a new tack—conditioning one sample point at a time runs into technical complications in general continuous spaces. The reader who is interested in seeing how this is done will find the general formulation of conditional expectation in the preamble to Problem XIII.13.

7 How well can we estimate?

Let X and Y be continuous random variables governed by some joint density. Given X what is the best estimate \tilde{Y} we can form of Y ? While the answer surely depends upon what we mean by “best”, a natural choice is based upon the idea of minimising the expected squared difference between Y and \tilde{Y} .

Suppose $g(X)$ is a new random variable obtained as a function of the coordinate variable X . (For the time being it will suffice if we imagine g to be piecewise continuous or otherwise sufficiently regular. Section XII.6 outlines what constitutes a permissible function.) Our goal then is to minimise $Q_g = E[(Y - g(X))^2]$ over all choices of g . While this appears a formidable task, we might anticipate that the conditional expectation of Y given X should, if

intuition is to serve at all, provide at least a principled estimator; we could do worse than to begin with it. Write $\Delta(X) = g(X) - \mathbb{E}(Y | X)$ in a nonce notation. Then $(Y - g(X))^2 = [(Y - \mathbb{E}(Y | X)) - \Delta(X)]^2$. If we expand the square and take expectations of both sides then, by linearity of expectation, we have

$$Q_g = \mathbb{E}[(Y - g(X))^2] = \mathbb{E}[(Y - \mathbb{E}(Y | X))^2] - 2\mathbb{E}[\Delta(X)(Y - \mathbb{E}(Y | X))] + \mathbb{E}[\Delta(X)^2].$$

The middle term on the right is daunting but, by conditioning on X , we find

$$\begin{aligned}\mathbb{E}[\Delta(X)(Y - \mathbb{E}(Y | X)) | X] &= \Delta(X)\mathbb{E}[(Y - \mathbb{E}(Y | X)) | X] \\ &= \Delta(X)[\mathbb{E}(Y | X) - \mathbb{E}(Y | X)] = 0\end{aligned}$$

as $\mathbb{E}(Y | X)$ and $\Delta(X)$ are both determined by X . (A naïve understanding of conditional expectation will suffice for this argument. The anxious reader will find the measure-theoretic i's dotted in Problems XIII.13–23 and XIV.46–51.) By (6.4) we see then that $\mathbb{E}[\Delta(X)(Y - \mathbb{E}(Y | X))] = 0$. (Observe the implicit appeal to total probability in removing the conditioning by taking expectation with respect to X .) Consequently, $Q_g = \mathbb{E}[(Y - \mathbb{E}(Y | X))^2] + \mathbb{E}[(\mathbb{E}(Y | X) - g(X))^2]$ (for every sensible g). The second term on the right is positive for any choice of g and is zero if, and only if, $g(X) = \mathbb{E}(Y | X)$, leading to a result as elegant as it is simple.

THEOREM 1 *If X and Y have finite variances, the conditional expectation $\mathbb{E}(Y | X)$ is the least-squares estimator of Y by X . Formally, $\mathbb{E}(Y | X) = \arg \min_g \mathbb{E}[(Y - g(X))^2]$.*

In practice, the utility of the result is limited by the fact that computation of the conditional expectation requires knowledge of the underlying joint distribution and it is natural to seek computationally simpler estimators which can be easily computed from data. Linear estimators are particularly tempting in this regard and we are led to formulate an alternative least-squares problem: *among all linear estimators $\tilde{Y} = aX + b$, which one minimises the expected squared difference $\mathbb{E}((Y - \tilde{Y})^2)$?* By centring and appropriate scaling we may as well suppose that X and Y each have zero mean and unit variance, and have covariance $\text{Cov}(X, Y) = \mathbb{E}(XY) = \rho$. By completing squares and exploiting linearity of expectation again, we obtain

$$\mathbb{E}[(Y - aX - b)^2] = \mathbb{E}(Y^2) + a^2\mathbb{E}(X^2) + b^2 - 2a\mathbb{E}(XY) - 2b\mathbb{E}(Y) + ab\mathbb{E}(X).$$

By rearranging terms we may write the expression on the right in the form $1 - \rho^2 + b^2 + (a - \rho)^2$ and it is now easy to see that $\mathbb{E}[(Y - aX - b)^2]$ is minimised by setting $a = \rho$ and $b = 0$.

THEOREM 2 *Suppose X and Y each have zero mean and unit variance, and let ρ be their covariance. Then $\rho X = \arg \min_{a,b} \mathbb{E}[(Y - (aX + b))^2]$ is the least-squares linear estimator of Y given X .*

By shifting and scaling, it is easy to express the result for non-centred random variables. Suppose $E(X) = \mu_1$, $E(Y) = \mu_2$, and suppose further that X and Y have strictly positive variances $\text{Var}(X) = \sigma_1^2$ and $\text{Var}(Y) = \sigma_2^2$. Let $\text{Cov}(X, Y) = \rho$. Our basic result applies to the centred and scaled variables $X^* = (X - \mu_1)/\sigma_1$ and $Y^* = (Y - \mu_2)/\sigma_2$, whence $\tilde{Y}^* = \rho X^*$ is the least-squares linear estimator of Y^* given X^* . By removing the centring and scaling, it follows that

$$\tilde{Y} = \frac{\rho\sigma_2}{\sigma_1}(X - \mu_1) + \mu_2 \quad (7.1)$$

is the least-squares linear estimator of Y given X .

Of course, the best linear estimator is, in general, suboptimal compared to the least-squares estimator. In this light, Example 6.3 reinforces the special nature of the normal distribution: *the least-squares estimator takes a linear form when the underlying distribution is normal.*

8 Galton on the heredity of height

Francis Galton was a Victorian polymath who brought a curious and inventive mind to bear on an eclectic range of interests. As a young man he made a name as an explorer, his pioneering cartographic survey of South West Africa earning him renown. (Based on his experiences he also authored a charming little book, *The Art of Travel; or, Shifts and Contrivances Available in Wild Countries*, for the adventurous Victorian, in which one finds vignettes like the following piece of advice when faced with a river crossing (Figure 3).

“In crossing a deep river, with a horse or other large animal, drive him in; or even lead him along a steep bank, and push him sideways, suddenly into the water; having fairly started him, jump in yourself, seize his tail, and let him tow you across.”

This advice was not universally popular.) On his return from his travels Galton plunged into a variety of investigations. He made notable contributions in meteorology (he prepared the first weather map and was the first to describe the anti-cyclone), founded the “London School” of experimental psychology, and campaigned strenuously for the directed improvement of the human stock, a programme he called “eugenics”. Following his half-cousin Darwin’s publication of the epochal *On the Origin of Species* in 1859, Galton was drawn irresistibly into questions on heredity and spent much of the rest of his life in considering questions of human variability, hereditary traits, and their implications to society. His wide-ranging investigations led him to collect prodigious amounts of data and then to invent new statistical methods to analyse them.

Galton made his Presidential Address to the British Association for the Advancement of Science in 1885 on the subject of “Types and their Inheritance”.



Figure 3: Galton weighs in on how to ford a deep river (from The Art of Travel).

Extensive earlier investigations on produce of seeds of different sizes but of the same species had led Galton to observe that

“... the offspring did not tend to resemble their parents in size, but always to be more mediocre than they—to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small.”

He concluded that the offspring exhibited a mean filial regression towards mediocrity (the mean of the population) and further that this was directly proportional to the parental deviation from it. In his 1885 lecture he sought to place this observation on an irrefutable mathematical basis based on a large sample of family records of human height that he had laboriously collected to articulate a “simple and far-reaching law” governing hereditary transmission. Normalising for differences in male and female heights, Galton contrasted the deviation of the reported heights of adult children from the mean value of height for the population (or “level of mediocrity”) with the corresponding deviation of parental heights (or, more accurately, the average of the heights of the two parents or “mid-parentage”) from the mean. He concluded that

“... we can define the law of regression very briefly. It is that the height-deviate of the offspring is, on the average, two-thirds of the height-deviate of its mid-parentage.”

Galton proceeded to explain this beguilingly simple law of regression through heredity, a devastatingly simple analogy providing ballast.

“The explanation of it is as follows. The child inherits partially from his parents, partly from his ancestry. Speaking generally, the further his genealogy

goes back, the more numerous and varied will his ancestry become, until they cease to differ from any equally numerous sample taken at hazard from the race at large. Their mean stature will then be the same as that of the race; in other words, it will be mediocre. Or, to put the same fact into another form, the most probable value of the mid-ancestral deviates in any remote generation is zero. ... The combination of the zero of the ancestry with the deviate of the mid-parentage is the combination of nothing with something, and the result resembles that of pouring a uniform proportion of pure water into a vessel of wine. It dilutes the wine to a constant fraction of its original alcoholic strength, whatever that strength may have been."

The data on which Galton based his conclusions are reproduced in Table 1.⁶ I

Heights of the Mid-parents in inches (#)	TABLE I. NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES												Medians	
	Heights of the Adult children (All Female heights have been multiplied by 1.08).													
	<	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>
> (5)	-	-	-	-	-	-	-	-	-	-	-	1	3	-
72.5 (6)	-	-	-	-	-	-	-	1	2	1	2	7	2	4
71.5 (11)	-	-	-	-	1	3	4	3	5	10	4	9	2	2
70.5 (22)	1	-	1	-	1	1	3	12	18	14	7	4	3	3
69.5 (41)	-	-	1	16	4	17	27	20	33	25	20	11	4	5
68.5 (49)	1	-	7	11	16	25	31	34	48	21	18	4	3	68.2
67.5 (33)	-	3	5	14	15	36	38	28	38	19	11	4	-	-
66.5 (20)	-	3	3	5	2	17	17	14	13	4	-	-	-	67.2
65.5 (12)	1	-	9	5	7	11	11	7	7	5	2	1	-	66.7
64.5 (5)	1	1	4	4	1	5	5	-	2	-	-	-	-	65.8
< (1)	1	-	2	4	1	2	2	1	1	-	-	-	-	-
Totals (205)	5	7	32	59	48	117	138	120	167	99	64	41	17	14
Medians	-	-	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	-	-

Table 1: Galton's height data.

have retained the original formatting of the table except, in the interests of compactness, for the elimination of a column of row sums (which the reader can readily adduce from the values provided) and the merging of mid-parentage heights and numbers into a single column. In the table, Galton listed the heights of 928 adult children with the heights of their mid-parentage, 205 in number, defined as the arithmetic average of the height of the two parents. All female heights were transmuted by a factor of 1.08 to obtain male equivalents (a procedure that Galton attributes as standard among anthropologists).

Galton's selection of bin granularity in which to pool his data was dictated at least partially by uncertainties in the modes of data collection which led him to conclude that his data were liable to an error of two-thirds of an inch or more. I have provided a visualisation of what the original data may have looked like in Figure 4 by jittering each of the mid-parent and children heights by a random value in the interval $(-0.5, 0.5)$. The reader who feels that this is

⁶F. Galton, "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute*, vol. 15, pp. 246–263, 1886. The height data appear in Table I, p. 248.

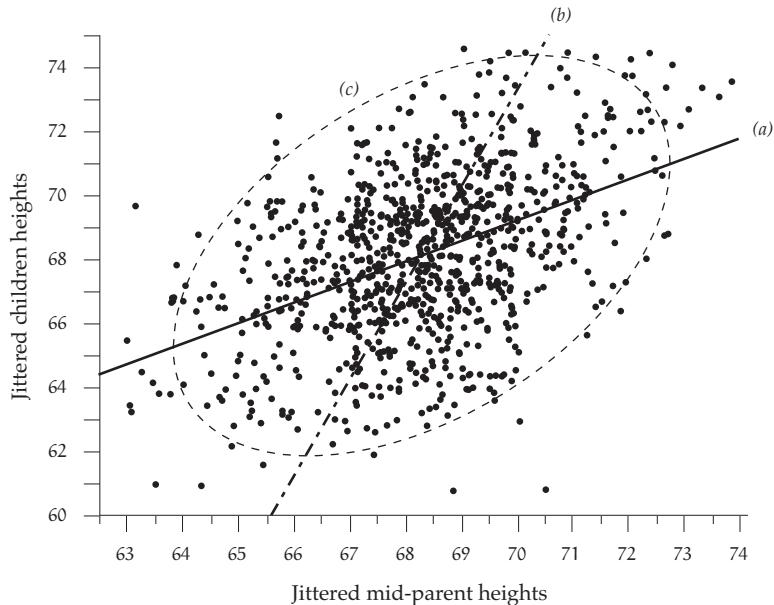


Figure 4: (a) An empirical linear least-squares estimator (solid line) of children's height from mid-parent heights gives the estimate $\hat{Y} = 24.32 + 0.64X$ for the jittered height data. (b) The corresponding linear least-squares estimator (dashed line) of mid-parent heights from children's height gives the estimate $\hat{X} = 46.10 + 0.33Y$. (c) The ellipse shown captures 95% of the area of a normal distribution with means, variances, and covariance matching the data.

a subjective fudge is invited to plot Galton's binned data directly. On examination of the figure, the reader may well agree with Galton's assessment that the statistical variations of height are extremely regular and generally conform to the abstract law of frequency of error, that is to say, the normal distribution. (The reader will find a principled test of this hypothesis in Section XX.11.) Under a normal hypothesis the least-squares estimate of children's height Y by mid-parent height X is linear and an empirical best-fit linear law can be readily extracted from the data. Galton reduced the tedium of hand calculation by fitting a line through the median values of children's height but it is no great chore today to work directly with a large mass of data.

The empirical linear least-squares estimator of children's height by mid-parent height minimises the arithmetic mean of the squared differences $(Y_j - aX_j - b)^2$ over the entire sample of 928 parent-child height pairs (X_j, Y_j) and yields the empirical version of (7.1) with the means, variances, and covariance estimated from the data sample. The slope of the empirical linear least-squares estimate of the jittered data shown by the solid line in Figure 4 matches Galton's estimate of $2/3$ and we appear to have independent confirmation of Galton's

observation that the filial generation regresses (statistically) towards the mean.

Galton had, however, fundamentally misread the data. If one reverses rôles, the linear least-squares estimate of mid-parent heights given children's heights shown by the dashed line in Figure 4 has a slope, not of $3/2$ as one may naïvely have expected, but actually of around $1/3$. Thus, we could argue on these grounds that the parental height was also regressing even faster to the mean. The apparent paradox is resolved by the fact that the least-squares estimate does not treat the two variables symmetrically—the estimate error is minimised in one direction only. What Galton had observed was the statistical artefact that remeasurement of an extreme-valued chance observation would, naturally, tend to move away from the extreme; a remeasurement of a value at the mean, however, would tend to show a symmetrical scatter around the mean value. It would appear hence that the remeasurements show a tendency to be less extreme than the original measurements though, in truth, the mean and variance are unchanged. What is true is that parent and children's heights are correlated. (The data support a correlation coefficient of around 0.5.) There is, however, no generational tendency to regression towards an ancestral mean.

Statistics owes the terms regression (regrettably) and correlation (happily) to Francis Galton. While some of his conclusions have since been repudiated, Galton had a far-reaching influence and inspired a new generation of statisticians, the redoubtable Karl Pearson among them. He was knighted for his contributions in 1909.

9 Rotation, shear, and polar transformations

Starting with a pair of random variables (X_1, X_2) drawn from a density $f(x_1, x_2)$ let us now consider the effect of a non-degenerate linear transformation of coordinates. Suppose

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{21}X_2, \\ Y_2 &= a_{12}X_1 + a_{22}X_2, \end{aligned} \tag{9.1}$$

where the matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is non-singular with determinant $J = a_{11}a_{22} - a_{12}a_{21} \neq 0$. As the reader is aware, from a geometric point of view the transformation accomplishes a rotation and a shear of the original coordinate system. An important special case is that of a pure rotation $A = \begin{pmatrix} \cos \vartheta & -\sin \vartheta \\ \sin \vartheta & \cos \vartheta \end{pmatrix}$.

Let $g(y_1, y_2)$ be the density of the pair (Y_1, Y_2) . To each sufficiently regular region \mathbb{I} in the new coordinate space (y_1, y_2) we may then associate the probability

$$P\{(Y_1, Y_2) \in \mathbb{I}\} = \iint_{\mathbb{I}} g(y_1, y_2) dy_2 dy_1.$$

On the other hand, (Y_1, Y_2) takes values in \mathbb{I} if, and only if, (X_1, X_2) take values in the preimage \mathbb{I}' consisting of the points (x_1, x_2) in the original coordinate

space that are mapped into \mathbb{I} by A . Writing $J = \det(A) = a_{11}a_{22} - a_{12}a_{21}$, it follows that

$$\begin{aligned}\mathbf{P}\{(Y_1, Y_2) \in \mathbb{I}\} &= \mathbf{P}\{(X_1, X_2) \in \mathbb{I}'\} = \iint_{\mathbb{I}'} f(x_1, x_2) dx_2 dx_1 \\ &= \iint_{\mathbb{I}} f\left(\frac{1}{J}a_{22}y_1 - \frac{1}{J}a_{21}y_2, -\frac{1}{J}a_{12}y_1 + \frac{1}{J}a_{11}y_2\right) \cdot \frac{1}{|J|} dy_2 dy_1,\end{aligned}$$

the final step following by the natural change of variable

$$\begin{aligned}x_1 &= \frac{1}{J}a_{22}y_1 - \frac{1}{J}a_{21}y_2, \\ x_2 &= -\frac{1}{J}a_{12}y_1 + \frac{1}{J}a_{11}y_2\end{aligned}$$

that is suggested by inverting the linear transformation A . The multiplicative factor $1/|J| = |\det(A^{-1})|$ in the integrand serves purely as a normalisation to compensate for the shearing of the axes introduced by A ; the reader will recognise it as the *Jacobian* of the transformation A^{-1} . Comparing integrands in the two integral expressions for the probability hence leads to the identity

$$g(y_1, y_2) = f\left(\frac{1}{J}a_{22}y_1 - \frac{1}{J}a_{21}y_2, -\frac{1}{J}a_{12}y_1 + \frac{1}{J}a_{11}y_2\right) \cdot \frac{1}{|J|}. \quad (9.2)$$

Linear transformations of normal variables provide fertile ground for inquiry.

EXAMPLES: 1) *Again, the bivariate normal.* Suppose $f(x_1, x_2) = \phi(x_1, x_2; \rho)$ is the bivariate normal density of (5.3). As it is obvious that the expression on the right of (9.2) yields a quadratic form in y_1 and y_2 in the exponent it follows that $g(y_1, y_2)$ is also a bivariate normal density. In consequence, *normality is preserved under linear transformations*.

2) *Coordinate rotations.* The linear map

$$(Y_1 \quad Y_2) = (X_1 \quad X_2) \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

corresponds to a 45° counter-clockwise rotation of coordinates. If (X_1, X_2) has the bivariate normal density $\phi(x_1, x_2; \rho)$, then simplification of (9.2) shows that the induced density of the pair of coordinates (Y_1, Y_2) in the rotated frame takes the form

$$g(y_1, y_2) = \frac{1}{\sqrt{2\pi(1+\rho)}} \exp\left(\frac{-y_1^2}{2(1+\rho)}\right) \cdot \frac{1}{\sqrt{2\pi(1-\rho)}} \exp\left(\frac{-y_2^2}{2(1-\rho)}\right).$$

It follows by inspection that Y_1 and Y_2 are independent, Y_1 has marginal normal density $\frac{1}{\sqrt{1+\rho}}\phi\left(\frac{y_1}{\sqrt{1+\rho}}\right)$, and Y_2 has marginal normal density $\frac{1}{\sqrt{1-\rho}}\phi\left(\frac{y_2}{\sqrt{1-\rho}}\right)$. A remarkable feature stands revealed: *there exists a rotation of coordinates for a bivariate normal which results in independent normal coordinates.* ►

As before, it is simpler to go to vector-matrix notation when there are more than two dimensions. Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has density $f(\mathbf{x}) = f(x_1, \dots, x_n)$. Consider a linear transformation on \mathbb{R}^n represented by the matrix of elements

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$$

and let $J = \det(A)$. With points in \mathbb{R}^n represented as *row* vectors per our convention, the map $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{XA}$ induces the density

$$g(\mathbf{y}) = f(\mathbf{y}A^{-1}) \cdot \frac{1}{|J|} \quad (9.2')$$

as the n -dimensional analogue of (9.2) in the new coordinate space for the random vector \mathbf{Y} ; the normative multiplicative factor $1/|J|$ again represents the Jacobian of the inverse transformation A^{-1} . The demonstration requires little more than replacing two-dimensional integrals by n -dimensional variants.

Similar considerations hold for other transformations, the familiar polar transformation $X_1 = R \cos \Theta, X_2 = R \sin \Theta$ being the most important of these. As the Jacobian of the map $(X_1, X_2) \mapsto (R, \Theta)$ is r , writing $g(r, \vartheta)$ for the density of the pair (R, Θ) , an entirely similar argument shows that

$$g(r, \vartheta) = f(r \cos \vartheta, r \sin \vartheta) \cdot r \quad \text{if } r > 0 \text{ and } -\pi < \vartheta \leq \pi.$$

The origin of the coordinate plane is left out of the specification but has probability zero and plays no rôle.

EXAMPLE 3) Polar coordinates. If (X_1, X_2) has density $(2\pi)^{-1} e^{-(x_1^2 + x_2^2)/2}$ then X_1 and X_2 are independent, each possessed of the standard normal marginal density $\phi(\cdot)$. The polar variables R and Θ then have joint density $g(r, \vartheta) = (2\pi)^{-1} r e^{-r^2/2}$ with $r > 0$ and $-\pi < \vartheta \leq \pi$. Integrating out r and ϑ separately shows that the marginal densities of R and Θ are $g_1(r) = r e^{-r^2/2}$ and $g_2(\vartheta) = (2\pi)^{-1}$ over $r > 0$ and $-\pi < \vartheta \leq \pi$, respectively, and it follows that R and Θ are independent random variables. ▶

A polar transformation in three dimensions replaces Cartesian variables x_1, x_2 , and x_3 by a radial distance $r > 0$, a longitude $-\pi < \varphi \leq \pi$, and a latitude $-\pi/2 < \vartheta < \pi/2$ via the transformations $X_1 = R \cos(\Theta) \cos(\Phi)$, $X_2 = R \cos(\Theta) \sin(\Phi)$, and $X_3 = R \sin(\Theta)$. The Jacobian of the transformation is $r \cos \vartheta$ and, accordingly, the joint density of (R, Θ, Φ) is given by

$$g(r, \vartheta, \varphi) = f(r \cos(\vartheta) \cos(\varphi), r \cos(\vartheta) \sin(\varphi), r \sin(\vartheta)) \cdot r \cos(\vartheta).$$

Again, the half-axes in the x_3 direction (corresponding to $\vartheta = \pm\pi/2$) are left out of the specification but have probability zero in any case.

10 Sums and products

Suppose X_1, \dots, X_n are independent random variables. The simplest, non-trivial random variable defined in terms of these basic coordinate variables is the random sum $S_n = X_1 + \dots + X_n$ and this plays a particularly important rôle. As before, for each i , let f_i represent the marginal density of X_i . Write $g_n(s)$ for the density of S_n . Determining g_n would require on the face of it a formidable integration over n dimensions. A *recursive procedure* has the advantage of reducing this daunting task to bite-sized pieces.

Begin with the sum $S_2 = X_1 + X_2$. The event $\{S_2 \leq s\}$ corresponds to the region in the plane defined by the inequality $X_1 + X_2 \leq s$. It follows that the distribution function of S_2 is given by

$$\begin{aligned} G_2(s) &= \mathbf{P}\{S_2 \leq s\} = \iint_{\{(x_1, x_2): x_1 + x_2 \leq s\}} f_1(x_1)f_2(x_2) dx_2 dx_1 \\ &= \int_{-\infty}^{\infty} f_1(x_1) \int_{-\infty}^{s-x_1} f_2(x_2) dx_2 dx_1 = \int_{-\infty}^{\infty} f_1(x_1) F_2(s - x_1) dx_1 \end{aligned}$$

where F_2 is the distribution function of X_2 . Differentiating both sides with respect to s hence shows that the density of S_2 is given by the integral equation⁷

$$g_2(s) = \int_{-\infty}^{\infty} f_1(x_1)f_2(s - x_1) dx_1. \quad (10.1)$$

The expression on the right is called the *convolution* of f_1 and f_2 and is denoted $f_1 * f_2(s)$. As the order of X_1 and X_2 may be interchanged without changing the sum S_2 , it follows that the roles of f_1 and f_2 may be interchanged and we may write

$$g_2(s) = \int_{-\infty}^{\infty} f_1(s - x_2)f_2(x_2) dx_2 \quad (10.1')$$

as an alternative expression to (10.1). Thus, $g_2(s) = f_1 * f_2(s) = f_2 * f_1(s)$ and convolution is commutative (as the reader may also verify by a simple change of variable inside the integral). The reader may wish to compare these integral expressions with the convolution sums that she encountered in the case of arithmetic random variables in Section 2.

It is difficult to overstate the importance of convolution in mathematics and the operation plays a critical rôle in probability. For our immediate purposes it suffices to observe that convolution is a smoothing operation which “spreads out” the marginal densities. This has important consequences as we will see in Chapter XVII.

⁷The picky reader may wish to consider why differentiation is permitted under the integral sign on the right.

A word of caution is in order here. While (10.1) asserts that if X_1 and X_2 are independent then the density of $X_1 + X_2$ is the convolution of the marginal densities of X_1 and X_2 , the converse statement is in general not true. In particular, it is possible to exhibit *dependent* random variables X_1 and X_2 the density of whose sum also perversely exhibits the convolution property (10.1). See Problems IX.9 and IX.10.

To return to the general problem, observe that for $n \geq 2$ we may write $S_n = S_{n-1} + X_n$. The key observation now is that *the random variables S_{n-1} and X_n are independent*. (Why?) It follows that the density of S_n is a convolution of the densities of S_{n-1} and X_n whence $g_n(s) = g_{n-1} * f_n(s)$. Induction very quickly shows that *the density of the sum S_n is given by the n-fold convolution $g_n(s) = f_1 * f_2 * \cdots * f_n(s)$* . As we've seen, convolution is a commutative operation so that the convolution integrals may be computed in any order.

In the special case of positive random variables the range of the convolution integral may be reduced: if X_1 and X_2 are independent, positive random variables then their sum $S_2 = X_1 + X_2$ has density

$$g_2(s) = \int_0^s f_1(x)f_2(s-x) dx.$$

Of course, when additional terms are added to the sum the expressions for the densities are modified appropriately.

As a final illustrative exercise in transformations, we consider products and ratios of continuous variables. These appear not infrequently in applications (though not nearly as commonly as sums) and will serve to illustrate calculations when coordinate variables are transformed into new variables.

Suppose X and Y are independent random variables with marginal densities $f(x)$ and $g(y)$, respectively, so that the pair (X, Y) has joint density $f(x)g(y)$. Now consider the derived random variable $Z = XY$ and let $H(z)$ and $h(z)$ denote the distribution function and density, respectively, of Z . The event $\{Z \leq z\}$ corresponds to the union of the regions $\{(x, y) : x \leq z/y, y > 0\}$ and $\{(x, y) : x \geq z/y, y < 0\}$ corresponding to the shaded region bounded by the hyperbolic asymptotes $xy = z$ shown in Figure 5. Integrating over the indicated

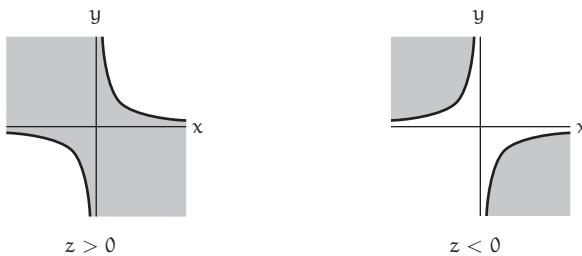


Figure 5: The hyperbolic regions defined by the equation $xy \leq z$.

region we obtain

$$\begin{aligned} H(z) &= \int_{-\infty}^0 dy g(y) \int_{z/y}^{\infty} dx f(x) + \int_0^{\infty} dy g(y) \int_{-\infty}^{z/y} dx f(x) \\ &= \int_{-\infty}^0 g(y) \left\{ 1 - F\left(\frac{z}{y}\right) \right\} dy + \int_0^{\infty} g(y) F\left(\frac{z}{y}\right) dy \end{aligned}$$

for every value of z . Formal differentiation shows that if $z \neq 0$ then

$$h(z) = - \int_{-\infty}^0 \frac{1}{y} g(y) f\left(\frac{z}{y}\right) dy + \int_0^{\infty} \frac{1}{y} g(y) f\left(\frac{z}{y}\right) dy = \int_{-\infty}^{\infty} \frac{1}{|y|} g(y) f\left(\frac{z}{y}\right) dy. \quad (10.2)$$

The density of $Z = XY$ may also be thought of as being arrived at by randomisation over y of Xy . For each y , the scaled variable Xy has (conditional) density $h(z | Y = y) = \frac{1}{|y|} f\left(\frac{z}{y}\right)$. As X and Y are independent, we rederive (10.2) by multiplying $h(z | Y = y)$ by the density $g(y)$ of Y and integrating out.

A similar exercise allows us to determine the density of the ratio $W = X/Y$ but we may finesse the calculations by expressing the ratio as a product. The reader has seen in Example 4.3 that the variable $T = 1/Y$ has density $\bar{g}(t) = t^{-2}g(t^{-1})$ and so, by (10.2), the variable $W = XT$ has density

$$\tilde{h}(w) = \int_{-\infty}^{\infty} \frac{1}{|t|} \bar{g}(t) f\left(\frac{z}{t}\right) dt = \int_{-\infty}^{\infty} |y| g(y) f(wy) dy, \quad (10.2')$$

the final step following by the natural change of variable $y = 1/t$ inside the integral.

11 Problems

1. *An expectation identity.* Suppose the distribution $p(k)$ of the arithmetic random variable X has support in the positive integers only. By rearranging the terms of the sum $\sum_{k \geq 0} kp(k)$, verify the useful identity $E(X) = \sum_{n=0}^{\infty} P\{X > n\}$.

2. *Urn problem.* An urn contains b blue and r red balls. Balls are removed at random until the first blue ball is drawn. Determine the expected number of balls drawn.

3. *Continuation.* The balls are replaced and then removed at random until the remaining balls are of the same colour. Find the expected number remaining in the urn.

4. *Pepys's problem, redux.* Is it more probable to obtain at least n sixes in $6n$ rolls of a die or to obtain at least $n+1$ sixes in $6(n+1)$ rolls of a die?

5. *Dice.* Let X and Y denote the number of ones and sixes, respectively, in n throws of a die. Determine $E(Y | X)$.

6. Show that it is not possible to weight two dice so that the sum of face values is equally likely to take any value from 2 through 12.

7. With ace counting as 1, extract the 20 cards with face values 1 through 5 from a standard 52-card deck and shuffle them. Ask a friend to think of a number N_1 between 1 and 5 but not to disclose it. Expose the cards one by one. Ask your friend to begin a count from the first card exposed and to take mental note of the face value, say, N_2 of the N_1 th card exposed; at which point she begins the count anew and takes note of the face value, say, N_3 of the N_2 th card from that point onwards (the $(N_1 + N_2)$ th card exposed overall); she proceeds in this fashion until all 20 cards are exposed at which point she, beginning with her private value N_1 , has accumulated a privately known sequence N_2, \dots, N_k of face values on which she “lands”. Her final noted face value, say, N_k for some value of k , corresponds to a card near the end of the deck with fewer than N_k cards left to be exposed. Propose a procedure to guess the value of N_k . (Remember that you don’t know her starting number N_1 .) Estimate by computer simulation your probability of success. [Hint: Do Problem I.16 first.]

8. *Reservoir sampling.* A finite reservoir (or *buffer*) at an internet router can store m packets of information. The first m incoming packets are accommodated. Thereafter, each incoming packet probabilistically replaces an existing packet in the reservoir. The protocol is as follows: for $j > m$, the j th arriving packet is accepted for placement in the reservoir with probability m/j and discarded with probability $1 - m/j$. If the packet is accepted then a packet in the reservoir is discarded at random and replaced by the incoming accepted packet. If a total of $n \geq m$ packets arrive at the reservoir show that each packet is accepted with probability m/n and discarded with probability $1 - m/n$, independently of the other packets. The beauty of the procedure is that it permits the selection of a random subset of m out of n entities *without prior knowledge of n*.⁸

9. For what value of C is the function $f(x) = C \exp(-x - e^{-x})$ a bona fide density? (This is the so-called *extreme value distribution*.) For what values of C and m is $f(x) = C(1 + x^2)^{-m}$ a density?

10. A particle of unit mass is split into two fragments of masses X and $1 - X$. The density f of X has support in the unit interval and by reasons of symmetry $f(x) = f(1-x)$. Let X_1 be the smaller of the two masses, X_2 the larger. The fragment of mass X_1 is split in like manner resulting in two fragments with masses X_{11} and X_{12} ; likewise, splitting X_2 results in masses X_{21} and X_{22} . We again use subscript 1 to denote the smaller of two masses, subscript 2 denoting the larger. Assume splits are independent, the split of a mass m governed by the appropriately scaled density $\frac{1}{m} f\left(\frac{x}{m}\right)$. Determine the joint density of X_{11} and X_{22} . Thence or otherwise determine the density of X_{11} .

11. Let $f(x, y) = ((1 + ax)(1 + ay) - a)e^{-x-y-\alpha xy}$ for $x, y > 0$. If $0 < a < 1$ show that f is a density.

12. *The triangular density.* Random variables X_1 and X_2 are independent, both uniformly distributed in the unit interval $(0, 1)$. Let $Y = X_1 + X_2$. Determine the density of Y , its mean, and its variance.

⁸The idea of reservoir sampling originated with C. T. Fan, M. E. Muller, and I. Rezucha, “Development of sampling plans by using sequential (item by item) selection techniques and digital computers”, *Journal of the American Statistical Association*, vol. 57, pp. 387–402, 1962. The idea can be put to use, for instance, to help defang Denial of Service attacks on an internet server: S. Khanna, S. S. Venkatesh, O. Fatemeh, F. Khan, and C. Gunter, “Adaptive selective verification: an efficient adaptive countermeasure to thwart DoS attacks”, *IEEE Transactions on Networking*, vol. 20, issue 3, pp. 715–728, June 2012.

13. The random variables X and Y have density $f(x, y) = 24xy$ with support only in the region $x \geq 0$, $y \geq 0$, and $x + y \leq 1$. Are X and Y independent?

14. The random variables X and Y are uniformly distributed in the annulus $a < \sqrt{x^2 + y^2} < b$. For each y , determine the conditional density of Y given that $X = x$.

15. Let X and Y be independent random variables with a common density. You are informed that this density has support only within the interval $[a, b]$ and that it is symmetric around $(a + b)/2$ (but you are not told anything more about the density). In addition, you are told that the sum $Z = X + Y$ has density $g_+(t)$ (which is given to you). How can you determine the density $g_-(t)$ of the variable $W = X - Y$ given this information?

16. *Additivity of expectation.* Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has density $f(x_1, \dots, x_n)$. Show by exploiting the linearity of integration that if the individual variables have expectation then $E(X_1 + X_2) = E(X_1) + E(X_2)$ and hence that $E(a_1 X_1 + \dots + a_n X_n) = a_1 E(X_1) + \dots + a_n E(X_n)$.

17. *Continuation, additivity of variance.* Suppose additionally that X_1, \dots, X_n are independent. First show that $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$ and thence that $\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$.

18. Show that $g(x, y) = e^{-x-y}/(x + y)$ is a density concentrated in the first quadrant $x > 0$, $y > 0$ in \mathbb{R}^2 . *Generalise:* Suppose f is a density concentrated on the positive half-line $(0, \infty)$. Then $g(x, y) = f(x + y)/(x + y)$ is a density in \mathbb{R}^2 . Find its covariance matrix.

19. Suppose X_1, \dots, X_n are independent variables with a common density f and d.f. F . Let $Y = \min\{X_1, \dots, X_n\}$ and $Z = \max\{X_1, \dots, X_n\}$. Determine the joint density of the pair (Y, Z) .

20. *Positive variables.* Suppose T is a positive random variable with d.f. $F(t)$ and associated density $f(t)$ with support in $[0, \infty)$. If T has mean μ and variance σ^2 , show that $\mu = \int_0^\infty (1 - F(t)) dt$ and $\sigma^2 + \mu^2 = 2 \int_0^\infty t(1 - F(t)) dt$. [Hint: Write $1 - F(t) = \int_t^\infty f(u) du$.]

21. *A waiting time problem.* Trains are scheduled on the hour, every hour, but are subject to delays of up to one hour. We model the delays T_j as independent random variables with a common d.f. $F(t)$ and associated density $f(t)$ with support in the unit interval $[0, 1]$. Let μ and σ^2 denote the mean and variance, respectively, of the delays. A passenger arrives at the station at a random time X . As, in units of one hour, only the fractional part of X matters we may suppose that X has been reduced modulo one and has the uniform density in the unit interval $[0, 1]$. (The examples of Section IX.4 provide some justification for the assumption of a uniform arrival density.) The time the passenger has to wait for a train to arrive is a positive variable W with density $g(t)$ and d.f. $G(t)$. Show that the conditional density of W given that $X = x$ is

$$g(t | X = x) = \begin{cases} f(t + x) & \text{if } 0 \leq t < 1 - x, \\ F(x)f(t + x - 1) & \text{if } 1 - x \leq t < 2 - x. \end{cases}$$

Show hence that $E(W) = \frac{1}{2} + \sigma^2$. Thus, the expected waiting time of the passenger increases proportionately with the variance of the delay. [Problem 20 comes in handy.]

22. *Continuation.* Determine $g(t)$. What is $G(1)$?

VIII

The Bernoulli Schema

Probability experiments with discrete outcomes shape much of our common intuition and feel for probability as a subject. Even in apparently simple settings, however, there are pitfalls for the unwary and unexpected sophistications. The most important of these cases deals with arithmetic distributions arising from the *Bernoulli schema*. The binomial and Poisson distributions that we have already encountered in limited scenarios are principal among these.

C 1, 2, 6, 7, 10
A 3–5, 9
F 8

1 Bernoulli trials

The simplest non-trivial arithmetic random variable arises in the modelling of a humble coin toss. Considerations do not get materially more difficult if we allow the coin to be biased in one direction or the other and accordingly we consider a “bent” coin whose probability of success (or “heads”) is p (where, of course, $0 \leq p \leq 1$).

One may wonder what the correspondence is between an abstract assignment of a success probability p to a “bent” coin and the actual performance of the experiment which can result in only one outcome. Roughly speaking, of course, what we mean is that if one performs very many “independent” tosses of the coin, then the fraction of tosses resulting in success will be approximately p . Thus, if $p = 1/4$, we expect roughly 5 out of 20 tosses to result in success. Our intuition leads us to expect somewhat better agreement with more tosses; with 1000 tosses we expect about 250 successes with a much smaller error rate. The link between the abstract theory and the experiment is furnished by the law of large numbers that we saw in Section V.6, where we explored coin tosses from a very different viewpoint.

Coin-tossing experiments of this form were the subject of investigation as long ago as the year 1713 when James Bernoulli’s influential work *Ars Conjectandi* was published posthumously. Probability experiments corresponding to a coin toss are hence called *Bernoulli trials* in homage to his trailblazing work.

Let us consider the toss of a bent coin with success probability p . It

will be convenient also to write $q = 1 - p$ for the probability of failure (or “tails”). If we imagine that instead of “heads” and “tails” the two coin faces show the numbers 1 and 0, respectively, then the outcome of a single coin toss is an arithmetic random variable X which takes the values 1 and 0 only with probabilities p and q , respectively. We then say that X is a *Bernoulli trial with success probability p*. If $p = 1/2$ we say that the Bernoulli trial is *symmetric*.

The simplicity of the distribution makes calculations trite. The expectation of X is $1 \cdot p + 0 \cdot q = p$ so that the mean of a Bernoulli trial is just the success probability itself. The variance of X is $(1 - p)^2 p + (0 - p)^2 q = q^2 p + p^2 q = pq(p + q) = pq$. Viewed as a function of p , the variance $pq = p(1 - p)$ is zero at the two endpoints $p = 0$ and $p = 1$, and is symmetric about $p = 1/2$ where it achieves its maximum value of $1/4$. Thus, there is most uncertainty in the outcome of a Bernoulli trial if the success probability is $1/2$ with the uncertainty in the outcome decreasing to zero as p approaches 0 or 1. Intuitive and natural.

We now consider a sequence X_1, \dots, X_n of independent Bernoulli trials, each with success probability p . (We may drop the word “independent” when dealing with Bernoulli trials; by convention, Bernoulli trials are assumed to be independent unless explicitly indicated otherwise.) The X_i then represent the outcomes of n independent tosses of a bent coin. Let $S_n = X_1 + \dots + X_n$. What can be said about the distribution of S_n ? The direct computation of an n -fold convolution sum does not appeal, even with the marginal distributions all Bernoulli. A recursive approach has the advantage of breaking the problem up into bite-sized pieces.

Let us adopt the nonce notation $b_n(k) = P\{S_n = k\}$ for the distribution of S_n . As $S_n = S_{n-1} + X_n$ and X_n takes values 0 and 1 only, the event $\{S_n = k\}$ occurs if, and only if, $S_{n-1} = k$ and $X_n = 0$, or $S_{n-1} = k - 1$ and $X_n = 1$. As S_{n-1} and X_n are independent, we hence obtain the recursive specification

$$b_n(k) = b_{n-1}(k)q + b_{n-1}(k-1)p. \quad (1.1)$$

While the recurrence is formally valid for values $n \geq 2$ and $k \geq 0$, it will be convenient to expand its range to all the integers. Accordingly, we define $b_n(k) = 0$ if n or k is < 0 , and set $b_0(0) = 1$. Then (1.1) is valid for all n and k excepting for $(n, k) = (0, 0)$. The reader should bear in mind that, for each $n \geq 0$, $b_n(k)$ is non-zero only for $0 \leq k \leq n$.

Starting with $n = 0$ we can now systematically build up the distribution $b_1(k)$ for $n = 1$, from which we can in turn deduce the distribution $b_2(k)$ for $n = 2$, and proceed iteratively in this fashion. The non-zero values of the distribution with the computation carried through $n = 5$ are shown arrayed in an inverse pyramid balancing on its apex in Table 1. The tip of the pyramid corresponds to $n = 0$ and each succeeding row increments n by 1. The n th row exhibits the values of $b_n(k)$ with k sweeping through the $n + 1$ integers from 0 to n sequentially. An examination of the pyramid of values should lead to a shrewd suspicion of the functional form of $b_n(k)$.

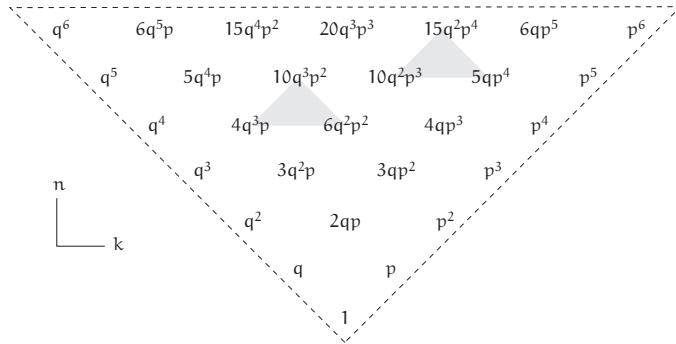


Table 1: The distribution of Bernoulli sums.

THEOREM Suppose X_1, \dots, X_n is a sequence of Bernoulli trials with success probability p . Then the sum $S_n = X_1 + \dots + X_n$ has distribution $b_n(k) = \binom{n}{k} p^k q^{n-k}$ with support only in $0 \leq k \leq n$.

We recall that, as a matter of convention, for real t and integer k , we define the binomial coefficients

$$\binom{t}{k} = \begin{cases} \frac{t(t-1)(t-2)\cdots(t-k+1)}{k!} & \text{if } k \geq 0, \\ 0 & \text{if } k < 0. \end{cases}$$

In particular, when $t = n$ is a positive integer then the coefficients $\binom{n}{k}$ are non-zero only for $0 \leq k \leq n$. (See Problems I.1–5.)

PROOF: As induction hypothesis, suppose that $b_{n-1}(k) = \binom{n-1}{k} p^k q^{n-1-k}$ for some n and all $0 \leq k \leq n-1$. Then

$$qb_{n-1}(k) + pb_{n-1}(k-1) = \left\{ \binom{n-1}{k} + \binom{n-1}{k-1} \right\} p^k q^{n-k} = \binom{n}{k} p^k q^{n-k},$$

the final step following via Pascal's triangle and completing the induction. ►

We have stumbled upon the august binomial distribution. Its importance and central rôle dictate a section all to itself.

2 The binomial distribution

We write $b_n(k) = b_n(k; p)$ to explicitly acknowledge the rôle of the parameter p . Then, as we have just seen,

$$b_n(k; p) = \binom{n}{k} p^k q^{n-k} \quad (k = 0, 1, \dots, n) \tag{2.1}$$

where, as before, $q = 1 - p$. (We have specified the range of k only to emphasise the fact that $b_n(k; p)$ has non-zero support only over the $n + 1$ integers $0 \leq k \leq n$.) An arithmetic random variable X with distribution $b_n(k; p)$ is said to have the *binomial distribution with parameters n and p* . The binomial is one of the triad of distributions of central importance in probability—the others are the Poisson and the normal distributions—and well repays attention.

The reader will have realised that we have made rather heavy weather over the derivation of the binomial distribution. If X has distribution $b_n(k; p)$ then we may identify X with the number of successes in n tosses of a bent coin with success probability p . Consider, for instance, a sample outcome with toss numbers i_1, \dots, i_k resulting in success and the remaining $n - k$ tosses resulting in failure. By independence of the tosses, this outcome has probability $p^k q^{n-k}$. As there are exactly $\binom{n}{k}$ ways of specifying the locations of k successes, additivity says that $P\{X = k\} = \binom{n}{k} p^k q^{n-k}$. And, of course, this is just the previously derived expression for $b_n(k; p)$. So was the effort in working recursively through the convolution relationship wasted? While the direct approach is much simpler, our earlier analysis had the salutary effect of exhibiting the convolution mechanism in all its glory in this simple case. The procedure pays dividends in more complex situations where direct combinatorial arguments will not, in general, be available.

While the binomial distribution is firmly associated with the number of successes in coin tosses it may be arrived at via other experiments, though these are all formally equivalent to coin tossing.

EXAMPLES: 1) *An urn model.* Suppose n balls are distributed in m urns. The probability that the first r urns receive exactly k balls is given by

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

If we set $p = r/m$, the latter expression is seen to be identical to $b_n(k; p)$. Thus if the probability of success is a rational number then we may model the number of successes in a coin-tossing experiment by an equivalent urn model. We can pass to irrational success probabilities p by considering the limit of a sequence of urn experiments with the numbers r/m converging to p .

2) *Guessing cards.* A card is drawn at random from a standard 52-card deck and a psychic attempts to guess its value. If sampling is with replacement and draws are independent, the probability that in n draws the psychic guesses correctly k times is $b_n(k; 1/52) = \binom{n}{k} \left(\frac{1}{52}\right)^k \left(\frac{51}{52}\right)^{n-k}$. Departures from this distribution may be viewed as evidence of supernatural insight. Or, for the sceptic, a rigged game.

3) *Pepys's problem.* Is it more probable to obtain at least one six in six rolls of a die or to obtain at least two sixes in twelve rolls of a die? Samuel Pepys famously

posed this problem to Isaac Newton in 1693—and disbelieved Newton's correct answer.¹ Let us make up our own minds on the score.

Obtaining a six on a given roll of a die may be modelled as the toss of a coin with success probability $1/6$. Accordingly, the probability of rolling no six in six rolls is $b_6(0; 1/6) = (5/6)^6 = 0.33\cdots$ while the probability of rolling fewer than two sixes in twelve rolls is $b_{12}(0; 1/6) + b_{12}(1; 1/6) = (5/6)^{12} + 12 \cdot (1/6)(5/6)^{11} = 0.38\cdots$. It is hence more probable to roll at least one six in six rolls than to roll at least two sixes in twelve rolls. ►

Suppose X has the binomial distribution $b_n(k) = b_n(k; p)$ where, for the rest of this section, we suppose that p is fixed and suppress it in the notation for clarity. It is easy to verify that $kb_n(k) = npb_{n-1}(k-1)$ for all k . It follows that $E(X) = \sum_k kb_n(k) = np \sum_k b_{n-1}(k-1)$. The sum on the right is equal to one as it sums all the terms in the binomial distribution $b_{n-1}(\cdot)$. It follows that *the mean of the binomial distribution $b_n(k; p)$ is np* . In particular, the expected number of successes in n tosses of a fair coin is $n/2$. Satisfactory.

A small combinatorial shimmy now allows us to write down the variance of X . Again utilising the identity $kb_n(k) = npb_{n-1}(k-1)$, we observe

$$k^2 b_n(k) = k[npb_{n-1}(k-1)] = np[(k-1)b_{n-1}(k-1) + b_{n-1}(k-1)].$$

By summing over k , it follows that

$$\sum_k k^2 b_n(k) = np \left[\sum_k (k-1)b_{n-1}(k-1) + \sum_k b_{n-1}(k-1) \right] = np[(n-1)p+1]$$

as we recognise the first sum in the intermediate step as just the mean of the binomial distribution $b_{n-1}(\cdot) = b_{n-1}(\cdot; p)$ while the second sum just adds over all the terms of the distribution. To finish up, $\text{Var}(X) = \sum_k k^2 b_n(k) - (np)^2$, whence *the binomial distribution $b_n(k; p)$ has variance npq* .

As a binomial variable may be written as a sum of Bernoulli trials, we could have obtained the mean and the variance directly by appeal to the additivity of expectation and variance but a direct verification has its points.

3 On the efficacy of polls

As the examples of the previous section indicate, we may in general associate the binomial distribution with experiments of the following form. Suppose A is any event in a probability space on which we perform n independent trials. If our only interest is in the occurrence or otherwise of A , we may associate with the trials the sequence of random variables Z_1, \dots, Z_n where, for each j , Z_j

¹For an account of the history see E. D. Schell, "Samuel Pepys, Isaac Newton, and probability", *The American Statistician*, vol. 14, pp. 27–30, 1960.

takes value 1 if the event A occurs in the jth trial and takes value 0 otherwise. In graphic terminology, the variables Z_j are called *indicator random variables* for the event A. It follows then that $P\{Z_j = 1\} = P(A)$ and $P\{Z_j = 0\} = 1 - P(A)$ or, in other words, Z_1, \dots, Z_n constitutes a sequence of Bernoulli trials with success probability $p = P(A)$. The sum $S_n = \sum_{j=1}^n Z_j$ then represents the number of occurrences of A in n independent trials and has the binomial distribution with parameters n and p. As far as the occurrence of A is concerned the possibly complex probability space of the original experiment is equivalent to repeated independent tosses of a bent coin with success probability p.

This explains the prevalence and importance of the binomial as many applications devolve around observations of a phenomenon in repeated independent trials. The knowledge of the distribution can be put to use in tests of hypotheses and evaluations of predictions.

EXAMPLES: 1) *Testing sera and vaccines.* Bacterial diseases were the scourge of livestock in the mid-twentieth century. Faced with the lethal bacterial enteritis known as “white scour”, for instance, the veterinarian had very few weapons of consequence in his armoury and would resort to sera such as *E. coli* antiserum which were widely considered to be ineffective.² Faced with repeated failures, the emergence of a new serum or vaccine would be greeted sceptically.

One way to judge the performance of a serum is to compare its performance on a test group with an unvaccinated control group. If the group given the serum exhibits performance similar to the unvaccinated group, this lends credence to the argument that the serum is worthless; contrariwise, if the test group exhibits markedly better performance than the control group, this can be taken as evidence in support of the effectiveness of the serum. For instance, suppose that 25% of cattle in a large population get infected. If a random sample of 10 cattle are inoculated with a test serum and none get infected, is that evidence of effectiveness? The probability that there are no infections in an unvaccinated control group of 10 cattle is $b_{10}(0; .25) = (.75)^{10} = 0.056\dots$. It is thus unlikely that the observed result occurs by chance and there are hence grounds to believe that the serum was effective. The general principle may be articulated as follows: *the more unlikely it is for claimed (positive) results for a serum to be duplicated in an unvaccinated control group, the greater the grounds for belief in the efficacy of the serum.* On these grounds, it is even stronger evidence for the effectiveness of a vaccine if two out of 23 vaccinated cattle become infected as the probability that two or fewer cattle get infected in an unvaccinated control group is only $b_{23}(0; .25) + b_{23}(1; .25) + b_{23}(2; .25) = 0.049\dots$.

2) *Confidence in polls.* A newspaper reports that the results of a poll show that 55% of the population support, say, a new health care initiative being bruited.

²Fortunately, the discovery of the sulphonanimides followed soon after by penicillin and the other wonder antibiotics in the middle of the twentieth century helped eradicate this scourge.

The report says that the poll has a margin of error of $\pm 3\%$ which, even in the worst case, would suggest that the majority support the new policy. How much faith should one place in such a pronouncement?

Any poll implements the following abstract idea. Select n individuals by independent sampling from the general population, an unknown fraction p of whom support the proposed policy (or initiative, individual, or party). (It is assumed individuals are equally likely to be selected; in the abstract experiment the sampling is with replacement though it makes little difference in practice if the population is large vis à vis the sample size.) Each of the n selected individuals tabulates their vote. Let Z_j be the indicator variable for the event that the j th individual votes in favour of the policy. Then $S_n = \sum_{j=1}^n Z_j$ is the number of individuals from the random sample who vote in favour and, as we've just seen, S_n has distribution $b_n(k; p)$.

How should one determine the unknown p ? Let us consider a *gedanken* experiment where the fraction of a population in support of the policy is a variable ζ which may vary continuously between the values of 0 and 1. With the sample size n fixed, the probability that for any given ζ a random sample of size n contains exactly S_n individuals in favour determines a function of ζ given by $b(\zeta) = b_n(S_n; \zeta) = \binom{n}{S_n} \zeta^{S_n} (1 - \zeta)^{n - S_n}$. Differentiation shows that $b(\zeta)$ achieves a unique maximum at the value $\zeta = \frac{1}{n} S_n$ and it is hence natural to estimate the unknown p in our test experiment by the statistic $\frac{1}{n} S_n = \arg \max_{\zeta} b_n(S_n; \zeta)$, called the *sample mean*, which we may identify as the mean of the *empirical distribution* which places equal mass $1/n$ on each of the sampled data points Z_1, \dots, Z_n . That is to say, we estimate p by the value that yields the largest probability of observations consistent with the data. This is the principle of *maximum likelihood* due to R. A. Fisher. Apodictic support for the principle is provided by the observation that $E(\frac{1}{n} S_n) = p$: we say hence that the sample mean is an *unbiased estimator* of p .

The sample mean $\frac{1}{n} S_n$ hence gives us a principled estimate of the unknown p from the given sample data. How confident are we that this is a good estimate? If we are willing to tolerate an error of $\pm \epsilon$ in the estimate, the probability that the estimate lies outside our tolerance of error is given by

$$\mathbf{P}\left\{\left|\frac{1}{n} S_n - p\right| \geq \epsilon\right\} = \mathbf{P}\{|S_n - np| \geq n\epsilon\} = \sum b_n(k; p) \quad (3.1)$$

where the sum on the right is over all k satisfying $|k - np| \geq n\epsilon$. If we write δ for the sum, we may expect the sample mean $\frac{1}{n} S_n$ to satisfy the inequalities $p - \epsilon \leq \frac{1}{n} S_n \leq p + \epsilon$, or, equivalently, $\frac{1}{n} S_n - \epsilon \leq p \leq \frac{1}{n} S_n + \epsilon$, with *confidence* $1 - \delta$. If δ is close to zero we have high confidence in our estimate of p by the sample mean; conversely, a value of δ near one leaves us with little or no confidence in the estimate.

The confidence parameter δ depends both on the sample size n and the unknown parameter p . It will suffice for our purposes, however, if we can show

that, for a given value of n , $\delta = \delta(n, p)$ may be *uniformly* bounded for all values of p . The device of Chebyshev that we encountered in Section V.6 points the way.

The sum on the right in (3.1) varies over all k satisfying the inequality $|k - np|/(\epsilon n) \geq 1$, and so we obtain the simple bound

$$\begin{aligned} \sum_{k:|k-np|\geq\epsilon n} b_n(k; p) &\leq \sum_{k:|k-np|\geq\epsilon n} \left(\frac{k - np}{n\epsilon} \right)^2 b_n(k; p) \\ &\leq \frac{1}{n^2\epsilon^2} \sum_{k=-\infty}^{\infty} (k - np)^2 b_n(k; p) = \frac{p(1-p)}{n\epsilon^2}, \end{aligned}$$

as allowing the index k to range over all integers in the penultimate step can only increase the value of the sum which we may now identify with the variance of the binomial. The expression on the right gives a bound for the confidence parameter δ but is still slightly unsatisfactory as it is couched in terms of the unknown p . We can obtain a uniform bound, however, via the observation that the function $f(p) = p(1-p)$ on the unit interval achieves its maximum value of $1/4$ when $p = 1/2$. We hence obtain the uniform bound

$$P\left\{ \left| \frac{1}{n} S_n - p \right| \geq \epsilon \right\} \leq \frac{1}{4n\epsilon^2}. \quad (3.1')$$

The right-hand side tends to zero as $n \rightarrow \infty$ uniformly in p and we hence have a validation of polling mechanisms: *the sample mean $\frac{1}{n} S_n$ provides a principled estimate of an unknown population proportion p in the sense that the error ϵ in the estimate is small with high confidence $1 - \delta$ provided the sample size is large enough.*

The utility of the bound on the right in (3.1') is not that it is sharp—it is not—but that it allows us to discern general features of the problem that are obscured by the algebraic detail. To provide a numerical illustration of the effect of the sample size on the confidence, suppose that 55% of the population actually does support policy, that is to say, $p = .55$. With $\epsilon = .03$ our confidence $1 - \delta$ that the sample mean $\frac{1}{n} S_n$ of a random sample is within 3% of $p = .55$ is given by summing the binomial probabilities $b_n(k; .55)$ over $.52n \leq k \leq .58n$. If the sample size is $n = 100$ a numerical evaluation shows that the confidence is a paltry 45% and the claimed result is questionable at best; the confidence improves to a still shaky 76% for $n = 350$; for $n = 1000$ it reaches 95% and at this level of confidence we may perhaps profess to have some faith in the claim.

3) *Sampling to determine the impact of an invasive species.* Environmental balance can be affected by the invasion of a foreign species introduced either inadvertently or purposefully. To pick just one instance, consider the case of a single exotic species of fish, the Nile Perch, which was introduced into the African

Great Lakes of Victoria, Malawi, and Tanganyika for purposes of subsistence and sports farming. An environmental disaster is in the making as the aggressive Nile Perch has rapidly established itself in the Lakes to the detriment of a great diversity of endemic species of cichlid fish that are native to the Lakes and are threatened with extinction. The foreign intruder now constitutes an unknown fraction p of the piscine population in the Lakes. How does one determine its impact and spread?

In a random sample of n fish caught in the area, let S_n represent the number of Nile Perch. Then S_n has distribution $b_n(k; p)$ and, emboldened by the law of large numbers, we estimate p by the sample mean $\frac{1}{n}S_n$. How large should the sample size be? The estimate (3.1') says that a sample size of around 5000 will suffice in order to be able to claim an error of no more than 3% with confidence at least 95%. The bound is a little wasteful; as we saw in the previous example, a sample size of $n = 1000$ suffices. In practice, generating a truly random sample of even this size presents its own problems.

4) *And so, what happened in the US Presidential Election of the year 2000?* In polling applications one requires not just an estimate of an event probability to within a fixed error margin, but a comparison between the probabilities of two opposed hypotheses. In this case it may not be possible to specify an admissible deviation from the mean independently of the underlying probabilities and if the two probabilities are close the entire sampling mechanism can break down.

A case in point is the bitterly contested US Presidential Election of the year 2000 between candidates George W. Bush and Albert Gore. The pivotal state in that election was Florida and the winner of that state would go on to claim the presidency. On the night of the election, news agencies taking exit polls³ first called the state for Gore, then for Bush, and finally retreated from the field in disarray as the fluctuating poll numbers made predictions untenable. Not surprisingly, given the official ballot counts that were eventually released: 2,912,790 ballots were ultimately deemed to have been cast for Bush while 2,912,253 were deemed to have been cast for Gore. If p denotes the fraction of the population of Florida voters for Bush and $q = 1 - p$ the fraction of the population for Gore then $p = .5000460\cdots$ and $q = .4999539\cdots$ ⁴. An estimate of p within 3% is now completely useless as what is required is not just p but a determination of whether p is larger than q . In this finely balanced situation, a 1% or even a .01% error margin would prove insufficient. For a conclusive prediction we would require an estimate of p (or q) accurate within $\epsilon = |p - q|/2$. For the Florida numbers this would require an error tolerance of less than .000046 \cdots . With a choice of $\delta = .05$ for a 95% confidence level, the

³Polls taken as voters exit voting booths.

⁴This ignores the votes for third-party candidate Ralph Nader. While his share of the vote was small it was potentially decisive if he drew votes primarily from one candidate.

sample size requisite by (3.1') is at least $n \geq 1/4(p - q)^2\delta^2$ which is in excess of 10^{10} . This is of course completely ridiculous. Better estimates of the binomial tail don't help either. The problem here is fundamental: *two opposed hypotheses with probabilities very close to each other can be separated only by an accurate person-by-person tabulation of the entire population.*

Unfortunately, counts of this size introduce their own errors, whatever the mechanism used, mechanical, electronic, optical—or human. And if the separation between candidates is very small, the intrinsic error in the counting mechanism will render any conclusion statistically meaningless. In such a situation all we can do is give up the entire business as a bad job—or ask the Supreme Court to intervene, in which case a new random element is introduced into the decision process. Any assertion of “truth” in this setting becomes a vacuous logical exercise but perhaps a necessary political reality. ►

Polls conducted in practice deviate from the abstract model in several ways. For one, it is quite difficult to generate a truly random sample; for instance, polling mechanisms tend to favour densely populated urban areas. Furthermore, a number of polled individuals will decline to participate with participation influenced by a complex admixture of factors including age, sex, race, location, and income; an expurgated sample comprised only of “a coalition of the willing” skews the statistics. Generating an unbiased sample of sufficient size hence becomes problematic and expensive. Studies also show that polling responses are materially affected by the wording of the questions; slight ambiguities in the interpretation of different formulations of a question can lead to dramatically different poll numbers.

The proper design of polls to take into account these and other factors approaches an art form and it is a sad fact that many polls are, not to put too fine a face on it, worthless. Factors such as these explain why it is possible in the same news report to read that one poll shows 55% support with a 3% error margin while another poll shows 45% support with a 2% error margin. In other words, *caveat emptor*: a poll without supporting data is worth, well, the paper it is reported on.

The next two sections may be read at the reader's discretion as an illustration of ideas related to the binomial.

4 The simple random walk

The binomial is intimately connected with the elementary theory of random walks. Starting at the origin of a one-dimensional system, a particle moves in a succession of independent steps where at each step it moves either one unit to the right with probability 1/2, or one unit to the left, also with probability 1/2. We say that the particle performs a *random walk on the line*. In more picturesque language, we could replace the particle by an inebriated customer performing

a random walk on a road after leaving a pub. Alternatively, consider a gambler with an unlimited line of credit who gambles in a succession of independent games and wins or loses one dollar each game, his net winnings (positive or negative) representing a random walk.

Suppose X_1, X_2, X_3, \dots is a sequence of independent arithmetic random variables with each X_j taking value $+1$ with probability $1/2$ and -1 with probability $1/2$. For each n , write $S_n = X_1 + \dots + X_n$; it will be convenient also to set $S_0 = 0$. The sequence of values S_n ($n \geq 0$) then represents successive positions of a random walk on the line. The walk can take values $S_0 = 0, S_1 = k_1, S_2 = k_2, \dots$ if, and only if, the sequence of integer values $\{k_j, j \geq 0\}$ with $k_0 = 0$ satisfies $|k_{j+1} - k_j| = 1$ for each $j \geq 0$. In such a case we call $(0 = k_0, k_1, k_2, \dots)$ a *sample path* of the walk. Each sample path uniquely determines the values of the X_j and, of course, vice versa.

The space of all possible sample paths is very rich. As the sequence $\{X_j\}$ is formally equivalent to a sequence of coin tosses, we recall that the sample points of the experiment may be formally identified with a continuum of points in the unit interval (Example I.7.7). For any given n , however, the values S_0, S_1, \dots, S_n specify a truncated segment of a sample path and are specified by the outcomes of n fair coin tosses (or, equivalently, a dyadic interval). In picturesque language, we say that the random walk *visits* k at step n if $S_n = k$; we say that a visit to $k \geq 0$ occurs through strictly positive values if the walk satisfies $S_1 > 0, \dots, S_{n-1} > 0, S_n = k$. Of particular interest are visits to 0 and in this case we say that the random walk *returns to the origin*.

A very useful geometric picture of the walk can be obtained by graphing the step numbers n along the x -axis and the walk positions S_n along the y -axis and interpolating linearly between successive walk positions as shown in Figure 1. Each sample path is then a continuous line comprised of piecewise-linear segments of slope ± 1 . The path gives the appearance of a sequence of mountain ridges (though some of the peaks may be underwater if the walk dips through negative values). If the walk visits $k \geq 0$ at step n through strictly positive values then the corresponding path will lie strictly above the x -axis through the first $(n-1)$ steps and pass through the point (n, k) .

Suppose $(0 = k_0, k_1, \dots, k_n)$ is the truncation of any sample path to n steps. Then $S_1 = k_1, \dots, S_n = k_n$ if, and only if, $X_1 = k_1 - k_0, \dots, X_n = k_n - k_{n-1}$ so that each truncated path corresponds to the specification of a particular set of values to X_1, \dots, X_n . Thus, there are exactly 2^n valid n -step paths and each of these has probability of occurrence exactly 2^{-n} . As all n -step paths have the same probability, to determine the probability of any event

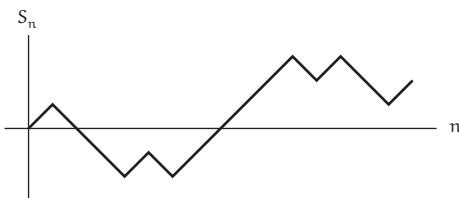


Figure 1: A sample path of a random walk.

defined by the first n steps it suffices to determine the number of paths that are favourable for the event.

For any positive integers n and k , let $N_n(k)$ denote the number of distinct paths from the coordinate origin $(0, 0)$ in the (x, y) plane to the point (n, k) , and let $N_n^+(k)$ denote the number of these paths that lie strictly above the x -axis for the first $(n - 1)$ steps. In this nomenclature, the probability that the walk visits k at step n is given by $P\{S_n = k\} = N_n(k)2^{-n}$. Likewise, the probability that the walk visits k at step n through strictly positive values is given by $P\{S_1 > 0, \dots, S_{n-1} > 0, S_n = k\} = N_n^+(k)2^{-n}$.

Before proceeding to the enumeration of paths, it will be convenient to adopt the temporary *convention* that $\binom{n}{t} = 0$ if t is not an integer to avoid the tedium of having to make repeated apologies for cases where t is an odd multiple of $1/2$. The enumeration of $N_n(k)$ is now a straightforward matter.

LEMMA 1 Suppose a, a', b, b' are integers and $0 \leq a < a'$. Then the number of paths from the point (a, b) in the plane to the point (a', b') depends only on the coordinate differences $a' - a = n$ and $b' - b = k$ and is given by $N_n(k) = \binom{n}{n+k}$.

PROOF: Suppose a path from (a, b) to (a', b') has r positive steps and s negative steps. Then $r + s = n$ and $r - s = k$ so that $r = (n + k)/2$ and $s = (n - k)/2$. As these positive and negative steps can be arranged in any order, the number of such paths is $\binom{n}{(n+k)/2}$, the case $(a, b) = (0, 0)$ and $(a', b') = (n, k)$ showing that this equals $N_n(k)$. ▶

The elegance of the geometric formulation becomes apparent in the famous *method of images* pioneered by D. André in 1887.

LEMMA 2 (THE METHOD OF IMAGES) Suppose $0 \leq a < a'$ and $b, b' > 0$ are integers. Then the number of paths from the point (a, b) to (a', b') that intersect the x -axis is equal to the total number of paths from the point $(a, -b)$ to (a', b') . This number depends only on $n = a' - a$ and $m = b' + b$ and is given by $N_n(m)$.

PROOF: Consider any path $(b = k_0, k_1, \dots, k_{n-1}, k_n = b')$ from (a, b) to (a', b') that intersects the x -axis. There must then be a *smallest* index j with $1 \leq j \leq n - 1$ for which $k_j = 0$, that is to say, $k_0 > 0, k_1 > 0, \dots, k_{j-1} > 0$, and $k_j = 0$. Let $A = (a, b)$ be the starting point of the path, $B = (a + j, 0)$ the point of first intersection with the x -axis, and $A' = (a', b')$ the endpoint of the path. If we reflect just the segment of the path from A to B below the axis, we obtain a mirror-image path from $A'' = (a, -b)$ to B as illustrated in Figure 2, and a corresponding path $(-b = -k_0, -k_1, \dots, -k_{j-1}, k_j = 0, k_{j+1}, \dots, k_{n-1}, k_n = y)$ from A'' to A' which intersects the x -axis for the first time at the point B . In this fashion each path from A to A' that intersects the x -axis may be put into a one-to-one correspondence with a path from A'' to A' . By the previous lemma, there are a total of $N_n(m)$ paths from $A'' = (a, -b)$ to $A' = (a', b')$. ▶

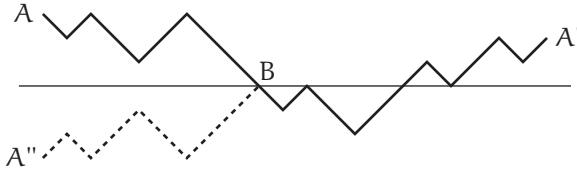


Figure 2: The method of images.

Suppose now that n and k are strictly positive integers. If a path progresses from $(0, 0)$ to (n, k) through strictly positive values only then its first step must be positive and it only remains to specify the remaining path from $(1, 1)$ to (n, k) . By Lemma 1, the total number of such paths is $N_{n-1}(k-1)$; of these, the number of paths that intersect the x -axis is $N_{n-1}(k+1)$ by Lemma 2. It follows that

$$N_n^+(k) = N_{n-1}(k-1) - N_{n-1}(k+1) = \binom{n-1}{\frac{n+k}{2}-1} - \binom{n-1}{\frac{n+k}{2}} = \frac{k}{n} \binom{n}{\frac{n+k}{2}}. \quad (4.1)$$

This result was first proved by W. A. Whitworth in 1878.

THE BALLOT THEOREM *The number of paths that progress from the origin to the point (n, k) through strictly positive values only is given by $N_n^+(k) = \frac{k}{n} N_n(k)$.*

The reader has seen this phenomenon explicated by very different methods in Example II.3.2 where the curious name of the theorem is explained.

5 The arc sine laws, will a random walk return?

Returns to the origin have the effect of restarting the random walk, the segments of a walk between returns to the origin being independent of one another. Now it is clear that a return to the origin can only occur on an even-numbered step and, in particular, the probability of a return to the origin at step $n = 2v$ is

$$u_{2v} := \mathbf{P}\{S_{2v} = 0\} = N_{2v}(0)2^{-2v} = \binom{2v}{v} 2^{-2v}. \quad (5.1)$$

What is the probability that the first return to the origin occurs *after* step $n = 2v$? The event $\{S_1 \neq 0, \dots, S_{2v} \neq 0\}$ occurs if either $\{S_1 > 0, \dots, S_{2v} > 0\}$ or $\{S_1 < 0, \dots, S_{2v} < 0\}$ occur, these two events having the same probability by symmetry. The occurrence of the event $\{S_1 > 0, \dots, S_{2v} > 0\}$ implies that $S_1 > 0, \dots, S_{2v-1} > 0$, and $S_{2v} = k$ for some even $k > 0$. Summing over all strictly positive, even k , we obtain

$$\begin{aligned} \mathbf{P}\{S_1 > 0, \dots, S_{2v} > 0\} &= 2^{-2v} \sum_k N_{2v}^+(k) \\ &\stackrel{(a)}{=} 2^{-2v} \sum_k [N_{2v-1}(k-1) - N_{2v-1}(k+1)] \stackrel{(b)}{=} 2^{-2v} N_{2v-1}(1) \stackrel{(c)}{=} \frac{1}{2} u_{2v} \end{aligned} \quad (5.2)$$

with (a) following by (4.1), (b) justified as the sum telescopes, and (c) following by Lemma 4.1 in consequence of the identity $N_{2v-1}(1) = \binom{2v-1}{v} = \frac{1}{2} \binom{2v}{v}$. We have obtained an intriguing result.

THE BASIC LEMMA *In a random walk the probability that the first return to the origin occurs after step $2v$ is equal to the probability of a return to the origin at step $2v$; in notation, $\mathbf{P}\{S_1 \neq 0, \dots, S_{2v} \neq 0\} = \mathbf{P}\{S_{2v} = 0\} = u_{2v}$ for each $v \geq 1$.*

This appears to have only curiosity value but has unexpected ramifications.

MAXIMA

With n fixed, where does the random walk achieve its maximum value? As the maximum value may be achieved repeatedly, let M_n denote the index m at which a walk S_0, S_1, \dots, S_n over n steps achieves its maximum value for the first time. Then, for any $0 < m < n$, the event $M_n = m$ occurs if, and only if, the conditions

$$\begin{aligned} S_m &> S_0, S_m > S_1, \dots, S_m > S_{m-1}, \\ S_m &\geq S_{m+1}, S_m \geq S_{m+2}, \dots, S_m \geq S_n \end{aligned} \quad (5.3)$$

all hold. The inequalities of the first row only depend on the variables X_1, X_2, \dots, X_m , while the inequalities of the second row only depend on $X_{m+1}, X_{m+2}, \dots, X_n$. Consequently, the sets of inequalities comprising each row represent two *independent* events.

As noticed by F. Pollaczek in 1952, a consideration of a *reversed walk* simplifies the inequalities in the first row to the setting of the basic lemma of this section. Set $X'_1 = X_m, X'_2 = X_{m-1}, \dots, X'_m = X_1$ and for $1 \leq k \leq m$ let $S'_k = X'_1 + \dots + X'_k$ denote the corresponding random walk. Then the inequalities in the first row of (5.3) are equivalent to $S'_m > 0, S'_{m-1} > 0, \dots, S'_1 > 0$. Resetting the walk after the first maximum, set $X''_1 = X_{m+1}, X''_2 = X_{m+2}, \dots, X''_{n-m} = X_n$ and let $S''_k = X''_1 + \dots + X''_k$ denote the corresponding random walk. The inequalities in the second row of (5.3) are then equivalent to $S''_1 \leq 0, S''_2 \leq 0, \dots, S''_{n-m} \leq 0$. The random walks $\{S'_j\}$ and $\{S''_j\}$ are independent and it follows that

$$\begin{aligned} \mathbf{P}\{M_n = m\} &= \mathbf{P}\{S'_1 > 0, \dots, S'_m > 0\} \cdot \mathbf{P}\{S''_1 \leq 0, \dots, S''_{n-m} \leq 0\} \\ &= \mathbf{P}\{S_1 > 0, \dots, S_m > 0\} \cdot \mathbf{P}\{S_1 \leq 0, \dots, S_{n-m} \leq 0\}, \end{aligned} \quad (5.4)$$

the final reduction following as, by symmetry, the walks $\{S_j\}$, $\{S'_j\}$, and $\{S''_j\}$ all have the same distribution. If $m = 2k$ or $2k+1$ the first term on the right is given by (5.2) to be equal to $\frac{1}{2} u_{2k}$. (The case when $m = 2k+1$ is odd follows because a return to the origin cannot occur on an odd-numbered step and so, if S_{2k} is strictly positive, then so also is S_{2k+1} .) The second term on the right

is, by the symmetry inherent in the situation, equal to the probability of the event $\{S_1 \geq 0, \dots, S_{n-m} \geq 0\}$. Now a strictly positive segment of a walk over $l+1$ steps, $S_0 = 0, S_1 > 0, \dots, S_l > 0, S_{l+1} > 0$, must have a first positive step. Resetting the origin to the point $(1, 1)$, the segment of the walk following the first step may now be placed in a one-to-one correspondence with a sample path of a walk over l steps which may return to the origin on occasion but never strays into strictly negative values. Accordingly,

$$\frac{1}{2} \mathbf{P}\{S_1 \geq 0, \dots, S_l \geq 0\} = \mathbf{P}\{S_1 > 0, \dots, S_{l+1} > 0\}.$$

We suppose for definiteness that $n = 2v$ and identify l with $n - m$. Grouping successive pairs of m values together for consideration, suppose $m = 2k$ or $m = 2k + 1$ for some $0 < k < v$. For each such k , we may now write (5.4) in the form

$$\mathbf{P}\{M_n = m\} = 2 \mathbf{P}\{S_1 > 0, \dots, S_m > 0\} \cdot \mathbf{P}\{S_1 > 0, \dots, S_{n-m+1} > 0\} = \frac{1}{2} u_{2k} u_{2v-2k} \quad (5.5)$$

by two applications of (5.2) to the probabilities on the right, the expression being the same whether $m = 2k$ or $m = 2k + 1$.

The cases $m = 0$ and $m = n$ are straightforward as $\mathbf{P}\{M_n = 0\} = \mathbf{P}\{S_1 < 0, \dots, S_n < 0\} = \frac{1}{2} u_{2v}$ and, likewise, if $M_n = n$, by considering the reversed walk, the origin becomes the point of first minimum and, by symmetry, it follows that $\mathbf{P}\{M_n = n\} = \frac{1}{2} u_{2v}$ as well.

The right-hand side of (5.5) is the product of two central terms of the binomial and the asymptotic estimate (VI.6.7) [or, (XIV.7.4)] applies to both provided k and $v - k$ are both large. For every fixed $0 < a < b < 1$, by summing over even and odd values $m = 2k$ and $m = 2k + 1$, we may hence estimate the probability that the last return to zero occurs between steps an and bn by

$$\mathbf{P}\{an \leq M_n \leq bn\} = \sum_{k=\lceil a\nu \rceil}^{\lfloor b\nu \rfloor} u_{2k} u_{2v-2k} \sim \sum_{\{k: a \leq \frac{k}{v} \leq b\}} \frac{1/v}{\pi \sqrt{\frac{k}{v}(1-\frac{k}{v})}} \quad (5.6)$$

asymptotically. In the sum on the right we recognise just the Riemann approximation to an integral; the limiting value of this sum obtained when $n \rightarrow \infty$ is hence given by

$$\int_a^b \frac{dx}{\pi \sqrt{x(1-x)}} = \frac{2}{\pi} \int_{\sqrt{a}}^{\sqrt{b}} \frac{du}{\sqrt{1-u^2}} = \frac{2}{\pi} \arcsin \sqrt{b} - \frac{2}{\pi} \arcsin \sqrt{a},$$

the change of variable $u^2 = x$ inside the first integral proving perspicacious. Thus, for each fixed choice of $0 < a < b < 1$, we have

$$\mathbf{P}\{an \leq M_n \leq bn\} \rightarrow \frac{2}{\pi} \arcsin \sqrt{b} - \frac{2}{\pi} \arcsin \sqrt{a}$$

as $n \rightarrow \infty$ and we stumble upon an unexpected limit law.

THEOREM 1 (ARC SINE LAW FOR MAXIMA) *For every $0 < t < 1$, the probability that the first maximum in a walk over n steps occurs on or before step tn satisfies $\mathbf{P}\{M_n \leq tn\} \rightarrow \frac{2}{\pi} \arcsin \sqrt{t}$ as $n \rightarrow \infty$.*

The graph of the limit law is shown in Figure 3. Of particular note is the rapid rise in the graph of the arc sine function near the points 0 and 1 and the slow rate of increase in the broad centre of the unit interval away from the endpoints. In consequence, as a surprise to native intuition, it is much more likely that a maximum of the segment of the random walk occurs near the beginning or the end of the segment than it is that a maximum occurs in the middle.

By symmetry, of course, an entirely similar arc sine limit law holds for the first appearance of a minimum of the segment of the walk.

The unexpected appearance of the arc sine law is due to the form of the expression on the right of (5.5). While the equation appears specialised to the setting of simple random walks, in reality it has a very general provenance. This connection with simple random walks explicates the startling arc sine laws discovered by Paul Lévy in the theory of fluctuations.⁵ We shall return to the theme in Section XV.8.

RETURNS TO THE ORIGIN

What is the probability that over a walk of $n = 2v$ steps, the last visit to the origin occurred at step $2k$? Let K_n denote the location of the last return to zero in a walk over n steps. (For definiteness, set $K_n = \infty$ if there is no return to zero.) We then have

$$\begin{aligned} \mathbf{P}\{K_n = 2k\} &= \mathbf{P}\{S_{2k} = 0, S_{2k+1} \neq 0, \dots, S_{2v} \neq 0\} \\ &\stackrel{(d)}{=} \mathbf{P}\{S_{2k} = 0\} \mathbf{P}\{S_{2k+1} \neq 0, \dots, S_{2v} \neq 0 \mid S_{2k} = 0\} = u_{2k} u_{2v-2k} \end{aligned}$$

where (d) follows because, conditioned on the event $\{S_{2k} = 0\}$, the walk begins anew from the origin at step $2k$ and the basic lemma applies to the next $2v - 2k$ steps. Summing over all k such that $a \leq k/n \leq b$ gives the same expression encountered on the right in (5.6) and in consequence we rediscover an arc sine law—for the last return to the origin this time.

THEOREM 2 (ARC SINE LAW FOR LAST VISITS) *For every fixed $0 < t < 1$, the probability that the last return to the origin in a walk over n steps occurs on or before step tn satisfies $\mathbf{P}\{K_n \leq tn\} \rightarrow \frac{2}{\pi} \arcsin \sqrt{t}$ as $n \rightarrow \infty$.*

⁵P. Lévy, "Sur certains processus stochastiques homogènes", *Composition Math.*, vol. 7, pp. 283–339, 1939.

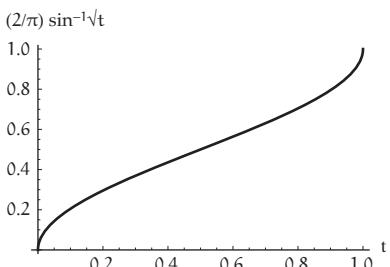


Figure 3: The arc sine limit law.

A sample path of a random walk (computer generated—the reader will readily understand that repetitively tossing a coin is an activity that quickly palls) over 5000 time steps is shown in Figure 4. Our example explains the surprising paucity of axis crossings: in view of the arc sine law for last visits, it is no longer

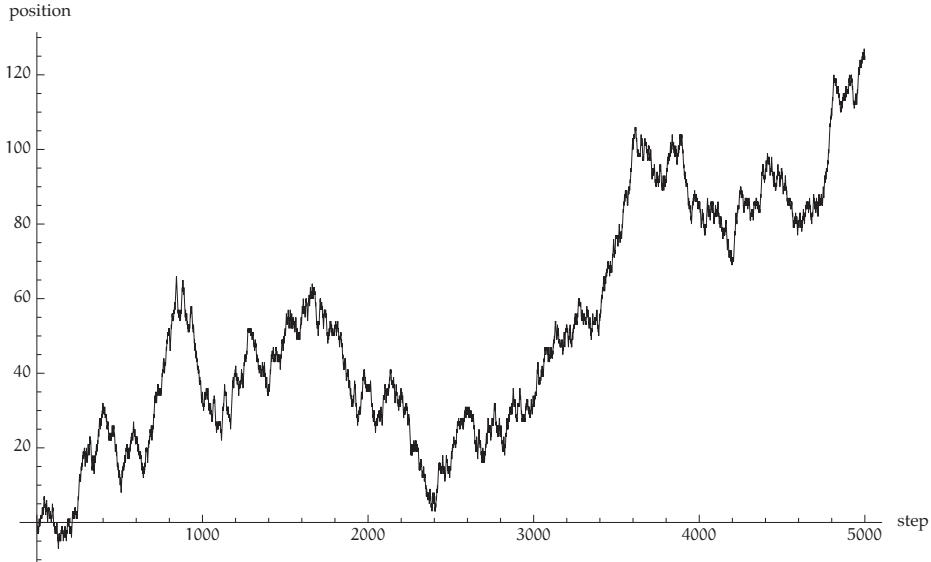


Figure 4: A sample path of a random walk.

surprising that the last return to zero occurs so early in the walk and we do not consider the walk shown to be anomalous.

EXAMPLE: *The tortoise and the hare.* Two players, say A and B, engage in a series of competitive events and a running total is kept of who is ahead. If over a sequence of 100 games player A gets into the lead early and never relinquishes it is that clear evidence of her superiority? Similar questions can be posed about, say, IQ tests: should Suzie be considered to be a stronger student than Johnny if she jumps out into the lead and stays there over a hundred tests? (The relevance of the ballot problem should be clear in this context.)

One can judge the statistical validity of the conclusion that A is the stronger player by considering a parallel experiment in which a fair coin is tossed repeatedly and a running total of the excess of successes over failures kept. Common intuition leads us to expect frequent changes in lead if the coin is fair. An observation hence that A is ahead of B over, say, the last half or the last two-thirds of the duration of the experiment may be taken in support of the premise that A is the superior player (that is, she is playing with a bent coin

favouring success). But intuition is not a reliable guide in itself and does not substitute for a careful mathematical analysis.

The probability that in a sequence of $2v$ fair coin tosses the last equalisation of heads and tails occurred at or before step $2tn$ with heads leading thereafter is given asymptotically by our theorem to be $B(t) = \frac{2}{\pi} \arcsin \sqrt{t}$. As $B(t)$ is symmetric about $t = 1/2$, the probability that one player or the other is in the lead over the last half of the game is $1/2$.

It is intuitively even more disturbing that the probability that one player stays in the lead through the last $(1-t)2v$ trials remains large even when t is quite small as can be seen from the graph of $B(t)$ near $t = 0$: the probability is approximately one-third that one player stays in the lead through the last three-quarters of the game; and about 20% of the time one player leads through the last 90% of the game. Thus, it is not at all unlikely that one player will lead the second through practically the entire duration of a game even when wins and losses are completely chance-dependent. ►

An additional conclusion may be gleaned from the arc sine law. Let q denote the probability that the walk *never* returns to the origin. If $K_{2v} \neq \infty$, there is a return to the origin in $2v$ steps and, accordingly, $1 - q \geq P\{K_{2v} \leq 2tv\}$ for any $0 < t \leq 1$ and any $v \geq 1$. By continuity of the arc sine function, for any $\epsilon > 0$, we have $1 - \epsilon < \frac{2}{\pi} \arcsin \sqrt{t} < 1$ for all t sufficiently close to 1. It follows that $q \leq P\{K_{2v} > 2tv\} \rightarrow 1 - \frac{2}{\pi} \arcsin \sqrt{t} < \epsilon$ for all sufficiently large v and t close to 1. As ϵ may be chosen arbitrarily small, this implies that $q = 0$ identically and the random walk will return to the origin with probability one. With each return to the origin, however, the situation resets and a new random walk progresses which ultimately returns to the origin and restarts the process anew. By induction there must be an infinity of returns to the origin.

THEOREM 3 *A random walk on the line returns to the origin infinitely often (i.o.)⁶ with probability one.*

A result which may explain the difficulty in weaning a drinker from the demon drink.

6 Law of small numbers, the Poisson distribution

What can be said when the success probability p of a binomial is small and the number of trials n is large and the two quantities are related in such a way that the expected number of successes $\lambda = np$ is moderate? Examples fitting such a paradigm could include misprints in a first folio edition of Shakespeare, gold dust in a California stream bank in 1849, electrons emitted at low intensity by

⁶This evocative notation and terminology due to Kai Lai Chung was introduced in Section IV.4.

k	0	1	2	3	4	5
Binomial	.064 346	.176 774	.242 578	.221 697	.151 808	.083 077
Poisson	.064 588	.176 954	.242 402	.221 372	.151 625	.083 082

Table 2: A tabulation of binomial and Poisson probabilities for $n = 1000$ and $p = 1/365$.

an electron gun, radioactive decay of an isotope, or Prussian soldiers killed in the year 1897 as a result of being kicked by their horses.

Under such conditions the binomial probability $b_n(k; p) = \binom{n}{k} p^k q^{n-k}$ may be approximated by $e^{-\lambda} \lambda^k / k!$, theoretical cover for the approximation being provided by Theorem IV.6.3, restated here for convenience.

POISSON'S APPROXIMATION TO THE BINOMIAL Suppose $np_n \rightarrow \lambda$, a fixed positive constant. Then $b_n(k; p_n) \rightarrow e^{-\lambda} \lambda^k / k!$ as $n \rightarrow \infty$ for every fixed positive integer k .

This result was proved by Siméon D. Poisson in 1837⁷ but its utility was only gradually perceived. Perhaps the first person to put Poisson's result to use in the statistical analysis of rare events was von Bortkewitsch, his analysis, published in 1898 under the beguiling title "The Law of Small Numbers", of the number of deaths of Prussian officers who were kicked by their steeds proving particularly memorable.⁸ The examples that follow may not have as sensational a character but arise from the same antecedent cause.

EXAMPLES: 1) *Birthdays*. An organisation has 1000 employees. What is the probability that exactly k of them have their birthdays on New Year's Day? We ignore leap-year complications and suppose a year has 365 days with birthdays uniformly distributed across them. The probabilities are then governed by a binomial with success probability $p = 1/365 = .00273 \dots$. The corresponding Poisson approximation has $\lambda = np = 1000/365$; these are listed in Table 2. As seen, for small values of k the Poisson approximation to the binomial is already good within two parts in 10,000.

2) *Poker*. As any poker aficionado knows, the probability of being dealt a hand of five cards with three face cards of the same value (three-of-a-kind) is quite small, approximately .021. In a marathon poker session of a hundred hands, the probability that a player gets 4 three-of-a-kind hands is then approximately $e^{-2.1} (2.1)^4 / 4!$ which is about .0992. The probability of getting 2 three-of-a-kind hands is much better, about .2700. The exact values given by the binomial are .0994 \dots and .2727 \dots , respectively.

⁷Siméon D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*. Bachelier, Paris, 1837.

⁸L. von Bortkewitsch, *Das Gesetz der kleinen Zahlen*. Teubner Verlag, Leipzig, 1898.

3) *Broadcast authentication.* Denial of service (which goes by the catchy sobriquet DoS) is a form of inimical attack on a communication medium in which an adversary attempts to overwhelm the medium by flooding it with fake messages. In a broadcast medium, for instance, the broadcaster can assure the receiver of the integrity of the message received by cryptographically signing key message packets. As cryptographic decoding is a computationally intensive process, an adversary who has access to sufficient bandwidth can attempt to swamp the receiver's computational resources by flooding the channel with packets containing fake signatures. The receiver will have to individually examine and eliminate these fake packets and in a DoS attack the large volume of fake packets can overwhelm her resources causing the connection to break.

A slightly more sophisticated strategy employed by the receiver can help defuse such attacks.⁹ The receiver randomly selects a fraction p of the purported signature packets for examination and discards the rest. The choice of the probability p is determined by the computational resources available to the receiver and the maximum expected adversarial traffic. If $p = .1$, for example, the receiver will discard 90% of the adversary's packets. Thus, most of the attack can be blunted at little cost. The danger, however, is that the true signed packet sent by the broadcaster is just as likely to be consigned to the discard heap as the fake packets. To mitigate against this possibility the broadcaster simply sends several copies of her signed packet. Suppose n copies are sent where $np = \lambda$ is moderate in value. The probability that at least one of the copies is examined by the receiver, thus validating the message stream, is then approximately $1 - e^{-\lambda}$. With $p = .1$ and $n = 25$ repetitions of the signed packet, the probability that at least one of the signed packets makes it through the blockade is approximately $1 - e^{-2.5} \approx .92$. The broadcaster can make assurance doubly sure by sending $n = 30$ copies of the signed packet; the receiver's confidence in acquiring at least one of these packets will now exceed 95%. And with 40 repetitions the confidence of acquisition exceeds 98%. Thus, the vast majority of adversarial packets can be rendered impotent even if the attacker is very able while, at extremely modest cost to the broadcaster, the receiver can be assured (with high confidence) of acquiring critical transmitted packets. ►

In view of their great importance, we resurrect the notation introduced in Section IV.6 for the limiting probabilities of Poisson's approximation and consider them on their own merits, divorced of their association with the binomial. For every fixed $\lambda > 0$, we set

$$p(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (k \geq 0).$$

⁹C. Gunter, S. Khanna, K. Tan, and S. S. Venkatesh, "DoS protection for reliably authenticated broadcast", *11th Annual Network and Distributed System Security Symposium*, San Diego, CA, February 2004.

A trite calculation shows that $\sum_{k=0}^{\infty} p(k; \lambda) = e^{-\lambda} \sum_{k=0}^{\infty} \lambda^k / k! = e^{-\lambda + \lambda} = 1$. It follows that $p(k; \lambda)$ determines an arithmetic distribution with support on the positive integers; the choice of the positive λ should be thought of as parametrising the distribution. The limiting values of Poisson's theorem hence determine a distribution in their own right; this is the important *Poisson distribution*. The Poisson, together with the binomial and the normal distributions, are the three principal distributions in probability with ramifications through the entire theory.

Let us now suppose that X is an arithmetic random variable conforming to a Poisson distribution with parameter λ ; that is to say, $P\{X = k\} = p(k; \lambda)$ for positive integers k . Manipulations similar to that for the binomial quickly turf up its expected value and variance.

Simple manipulations show that $kp(k; \lambda) = \lambda p(k - 1; \lambda)$ for all k . Summing over k it follows that $E(X) = \lambda$ as the sum of the terms $p(k - 1; \lambda)$ as k varies over the integers is identically one. We may thus identify the parameter λ with the mean of X ; it is hence permissible and indeed usual to say that a random variable X possessed of the distribution $p(k; \lambda)$ is Poisson with mean λ .

Suppressing the fixed parameter λ and writing $p(k; \lambda) = p(k)$ for compactness, we likewise observe that

$$k^2 p(k) = \lambda k p(k - 1) = \lambda(k - 1)p(k - 1) + \lambda p(k - 1) = \lambda^2 p(k - 2) + \lambda p(k - 1).$$

By summing over k we hence obtain $\text{Var}(X) = \sum_k k^2 p(k) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$. As a mild curiosity we note that the mean and the variance of X are both λ .

Now let us suppose X_1 and X_2 are independent Poisson random variables with means λ_1 and λ_2 , respectively. The sum $S_2 = X_1 + X_2$ then has distribution given by the convolution

$$\begin{aligned} P\{S_2 = k\} &= \sum_j p(j; \lambda_1) p(k - j; \lambda_2) = \sum_{j=0}^k e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_1^j \lambda_2^{k-j}}{j!(k-j)!} \\ &= e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_2^k}{k!} \sum_{j=0}^k \binom{k}{j} \left(\frac{\lambda_1}{\lambda_2}\right)^j = e^{-(\lambda_1 + \lambda_2)} \frac{\lambda_2^k}{k!} \left(1 + \frac{\lambda_1}{\lambda_2}\right)^k \end{aligned}$$

by the binomial theorem. It follows that

$$P\{S_2 = k\} = e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^k}{k!} = p(k; \lambda_1 + \lambda_2)$$

so that S_2 is Poisson with the sum mean $\lambda_1 + \lambda_2$. Induction quickly yields a strong stability property.

THEOREM Suppose $\{X_i, i \geq 1\}$ is a sequence of independent random variables where each X_i is Poisson with mean λ_i . Then, for each n , the sum $S_n = X_1 + \dots + X_n$ is

Poisson with mean $\lambda_1 + \dots + \lambda_n$ and, a fortiori, the family of Poisson distributions $\{p(\cdot; \lambda), \lambda > 0\}$ is closed under convolutions.

Thus, the Poisson distribution is stable under convolutions. I shall close this section with an example illustrating a combination of the binomial and the Poisson distributions.

EXAMPLE 4) *Chromosome breakage and repair.* Suppose that over a period of time the number of chromosomes that break is governed by a Poisson distribution with mean λ . A broken chromosome is repaired with probability p , the repair independent of similar efforts at healing other broken chromosomes and the total number of broken chromosomes. Let B be the number of broken chromosomes and R the number of these that are repaired. Then B has distribution $p(\cdot; \lambda)$ and R has the conditional binomial distribution $b_n(\cdot; p)$ given that $B = n$. To determine the unconditional distribution of R , condition on B and sum over all possibilities. With $q = 1 - p$ as usual, additivity yields

$$\begin{aligned} P\{R = k\} &= \sum_n P\{R = k | B = n\} P\{B = n\} = \sum_{n=k}^{\infty} \binom{n}{k} p^k q^{n-k} e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \frac{e^{-\lambda} (\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{(\lambda q)^{n-k}}{(n-k)!} = e^{-\lambda + \lambda q} \frac{(\lambda p)^k}{k!} = e^{-\lambda p} \frac{(\lambda p)^k}{k!}. \end{aligned}$$

Thus R has distribution $p(k; \lambda p)$ and the expected number of repaired chromosomes is λp . On consideration the reader may feel that this is reasonable—the expected number of broken chromosomes is λ and the expected fraction of these that are repaired is p . ▶

Observations fitting the Poisson distribution are legion. Examples in addition to those aforementioned include the distribution of organisms in a petri dish, customer arrivals to a queue, the distribution of dandelion seeds, stars in space, telephone connections to a wrong number, chromosome interchanges in cells, and even the distribution of flying bomb hits in London during the Second World War. The importance of the distribution stems both from its intimate connection to the binomial and its native application to problems in natural domains.

7 Waiting time distributions

Let us consider a succession of Bernoulli trials with success probability $p > 0$ (and corresponding failure probability $q = 1 - p < 1$). Let W denote the number of trials before the first success is obtained. Then W connotes the *waiting time to the first success*. Clearly, $W = k$ if, and only if, there are k consecutive failures followed by a success from which it follows that

$$w(k; p) := P\{W = k\} = q^k p \quad (k \geq 0).$$

Now $\sum_k w(k; p) = p \sum_{k=0}^{\infty} q^k = p/(1-q) = 1$ as the geometric series converges whenever $p > 0$. Thus, $w(k; p)$ is a bona fide distribution for all $p > 0$. For natural reasons we call $w(\cdot; p)$ the *geometric distribution with parameter p*.

The mean and variance of the distribution $w(k; p)$ require the computation of series of the form $\sum_{k \geq 0} kq^k$ and $\sum_{k \geq 0} k^2 q^k$ (both series being easily seen to be convergent by, say, the integral test). These sums are reminiscent of geometric series and indeed fall out quickly by elementary manipulations of the familiar series identity $\sum_{k \geq 0} q^k = 1/(1 - q)$ which is valid for all $|q| < 1$. For $K \geq 1$ it will be convenient to begin with a consideration of the sum $S_K = \sum_{k \geq K} kq^k$. As the terms of the series are positive we may rearrange terms and write it in the form

$$S_K = K(q^K + q^{K+1} + q^{K+2} + \dots) + q^K(q + 2q^2 + 3q^3 + \dots).$$

Factoring out q^K in the first term on the right leaves the familiar geometric series and we hence obtain

$$S_K = \frac{Kq^K}{1-q} + q^K S_1 \quad (K \geq 1). \quad (7.1)$$

In particular, when $K = 1$ we have $S_1 = \frac{q}{1-q} + qS_1$ and so, solving for S_1 , we get $S_1 = q/(1-q)^2$. By substitution in (7.1) we obtain the explicit solution

$$S_K = \frac{Kq^K}{1-q} + \frac{q^{K+1}}{(1-q)^2} \quad (K \geq 1).$$

The first of the two sums to be evaluated may be identified with S_1 and so no further work needs to be done to evaluate it:

$$\sum_{k \geq 0} kq^k = q + 2q^2 + 3q^3 + \dots = \frac{q}{(1-q)^2}.$$

Rearranging terms in the second of the sums to be evaluated (again permissible as the summands are positive), we now have

$$\begin{aligned}\sum_{k \geq 0} k^2 q^k &= \sum_{k \geq 0} k q^k + \sum_{k \geq 0} (k-1) k q^k \\&= (q + 2q^2 + 3q^3 + \dots) + (1 \cdot 2q^2 + 2 \cdot 3q^3 + 3 \cdot 4q^4 + \dots).\end{aligned}$$

We recognise the first sum on the right as just S_1 while the second sum on the right may be written via another rearrangement of terms in the form

$$\begin{aligned}
 2q^2 &+ 3q^3 + 4q^4 + 5q^5 + \dots \\
 &+ 3q^3 + 4q^4 + 5q^5 + \dots \\
 &\quad + 4q^4 + 5q^5 + \dots \\
 &\quad + 5q^5 + \dots
 \end{aligned}$$

where, with the index $K = 2, 3, \dots$ running through the rows of the array, we recognise in the K th row sum just the term S_K . It follows that

$$\begin{aligned} \sum_{k \geq 0} k^2 q^k &= S_1 + S_2 + S_3 + \dots = \sum_{K \geq 1} \frac{K q^K}{1 - q} + \sum_{K \geq 1} \frac{q^{K+1}}{(1 - q)^2} \\ &= \frac{S_1}{1 - q} + \frac{q^2}{(1 - q)^3} = \frac{q + q^2}{(1 - q)^3}. \end{aligned}$$

With these preliminary calculations out of the way, we may now simply read out the mean and the variance of the distribution $w(k; p)$. As $1 - q = p$, the mean is given by $E(W) = \sum_{k \geq 0} k q^k p = q/p$, while for the variance we have $\text{Var}(W) = \sum_{k \geq 0} k^2 q^k p - (q/p)^2 = q/p^2$. Another elementary approach via differentiation is sketched in the *Problems* at the end of this chapter.

EXAMPLES: 1) *The fair value of a bet.* In a sidewalk game two dice are rolled at a dollar per roll with the player winning if she rolls a double six. What odds should the sidewalk casino offer if the game is fair?

Let X denote the number of rolls (i.e., the player's dollar cost) before winning once. Then X conforms to a geometric distribution shifted by one unit, $p_k = P\{X = k\} = q^{k-1} p$, where $p = 1/36$ and $q = 1 - p = 35/36$. It follows that $E(X) = \sum_{k=1}^{\infty} k q^{k-1} p = 1/p = 36$. Offered odds of 36 for 1 would hence be *fair in the classical sense*.

2) *Stop-and-wait packet transmission protocols.* In a digital communication link messages are usually broken up into packets of information which are transmitted sequentially. Link protocols over such links are charged with ensuring reliable recovery of all transmitted packets in sequence while faced with a noisy or unreliable medium in which transmitted packets may be lost or corrupted beyond repair. One of the simplest mechanisms deployed relies upon feedback from the receiver: in its simplest form the transmitter transmits the currently expected packet and then waits for a response from the receiver; if the packet is received error-free the receiver sends back a short positive acknowledgement or ACK; if, however, the packet is received with unrecoverable errors, the receiver sends back a short negative acknowledgement or NAK. If the transmitter receives an ACK, she then transmits the next packet in sequence. If she receives a NAK on the other hand (or a time equal to the round-trip transmission delay elapses without receiving feedback from the receiver) she assumes that the packet is either lost or irretrievably corrupted and retransmits the packet.

If each packet is corrupted or lost independently of the others with probability $q = 1 - p$ and we assume ACKs and NAKs are always received error-free (a reasonable assumption in many situations as acknowledgements may be considered to be very short packets which are much less vulnerable

than the much longer data packets that are transmitted) then the number of re-transmissions of a given packet before successful receipt is a geometric random variable with parameter p . Taking into account the final successful transmission, the mean overhead involved in the retransmissions is $1 + q/p = 1/p$.

3) *The coupon collector's problem, revisited.* A cereal manufacturer includes one of c distinct coupons in each box of cereal. A coupon collector systematically buys boxes of cereal until he has collected all the coupons. Suppose that any given box of cereal is equally likely to contain any of the c coupons. Then the number of purchases between the acquisitions of the j th and $(j+1)$ th distinct coupons is geometrically distributed with parameter $p = (c-j)/c$. The mean number of purchases between the j th and $(j+1)$ th acquisitions is $q/p = j/(c-j)$. ►

A consideration of the geometric tail reveals a remarkable feature. Observe first that $P\{W \geq k\} = q^k$ as the waiting time to the first success is at least k if, and only if, the first k trials result in failure; of course, it comes to the same thing to add up the terms in the tail of the distribution. (Specialising to the symmetric case $p = q = 1/2$ yields the interesting result $P\{W > k\} = P\{W = k\} = 2^{-k-1}$.) Suppose now that an observer reports that the first j trials have resulted in consecutive failures. What is the conditional probability that an additional k trials will elapse without recording a success? As the occurrence of the event $\{W \geq j+k\}$ manifestly implies the occurrence of $\{W \geq j\}$, we obtain

$$P\{W \geq j+k \mid W \geq j\} = \frac{P\{W \geq j+k, W \geq j\}}{P\{W \geq j\}} = \frac{P\{W \geq j+k\}}{P\{W \geq j\}} = \frac{q^{j+k}}{q^j} = q^k.$$

On the right we recognise just the probability of the event $\{W \geq k\}$. Remarkably, for every pair of positive integers j and k , we find that

$$P\{W \geq j+k \mid W \geq j\} = P\{W \geq k\}, \quad (7.2)$$

which we may also write in the equivalent form

$$P\{W \geq j+k\} = P\{W \geq j\} P\{W \geq k\}. \quad (7.2')$$

In words, the waiting time to success given that a success has not yet occurred is independent of past history.

Arithmetic distributions satisfying (7.2) are said to have the *memoryless property* as they are oblivious to past history. Thus, geometric distributions have the memoryless property, but more can be said: *a distribution with support on the positive integers has the memoryless property if, and only if, it is geometric.*

The proof of the proposition is elementary but is essentially analytic and may be skipped without loss of continuity. Suppose W is any random variable taking values in the positive integers. Let $Q(k) = P\{W \geq k\}$ denote the right tail of the distribution of W . Then W has a memoryless distribution if, and only if, $Q(j+k) = Q(j)Q(k)$ for all

positive integers j and k . In particular, $Q^2(0) = Q(0)$ whence $Q(0)$ must be either 0 or 1. As the distribution has support in the positive integers it follows that $Q(0) = 1$. Suppose $Q(1) = x$. By induction it follows quickly that $Q(k) = Q(k-1)Q(1) = Q^k(1) = x^k$ and in consequence $P\{W = k\} = P\{W \geq k\} - P\{W \geq k+1\} = x^k - x^{k+1} = x^k(1-x)$ for all positive integers k . Thus, any memoryless distribution with support in the positive integers must necessarily be geometric.

Suppose W_1 and W_2 are independent geometric random variables with common success parameter p . Suppose $T_2 = W_1 + W_2$ and let $w_2(k; p) = P\{T_2 = k\}$ be its distribution. Then

$$w_2(k; p) = \sum_{i=0}^k w(i; p)w(k-i; p) = \sum_{i=0}^k (q^i p)(q^{k-i} p) = (k+1)q^k p^2 \quad (k \geq 0).$$

Working along these lines, a little diligence allows us to guess the general form of the distribution of a sum of independent geometric variables.

THEOREM Suppose $\{W_i, i \geq 1\}$ is a sequence of independent geometric random variables with common success parameter $p > 0$. For each r , let $T_r = W_1 + \dots + W_r$ and let $w_r(k; p) = P\{T_r = k\}$ be its distribution. Then $w_r(k; p) = \binom{r+k-1}{k} q^k p^r$.

PROOF: As $T_r = T_{r-1} + W_r$ is the sum of two independent random variables, we have $w_r = w * w_{r-1}$. Suppose $w_{r-1}(i; p) = \binom{r-1+i-1}{i} q^i p^{r-1}$ as induction hypothesis. Writing out the convolution sum, by Pascal's triangle we have

$$w_r(k; p) = q^k p^r \sum_{i=0}^k \binom{r+i-2}{i} = q^k p^r \sum_{i=0}^k \left[\binom{r+i-1}{i} - \binom{r+i-2}{i-1} \right].$$

The sum on the right telescopes to $\binom{r+k-1}{k} q^k p^r$, completing the induction. ►

In view of the identity $\binom{r+k-1}{k} = (-1)^k \binom{-r}{k}$, we may also write the distribution of T_r in the form

$$w_r(k; p) = \binom{r+k-1}{k} q^k p^r = \binom{-r}{k} (-q)^k p^r. \quad (7.3)$$

In consequence, the distribution $w_r(\cdot; p)$ is called the *negative binomial distribution with parameters r and p*. By definition of the binomial coefficients, it is clear that the negative binomial distribution $w_r(\cdot; p)$ has support only over the positive integers. Using the same artifice we've seen for the geometric distribution it is now easy to determine the mean, $E(T_r) = rq/p$, and variance, $Var(T_r) = rq/p^2$. These are just r times the mean and the variance, respectively, of the corresponding geometric distribution as is confirmed by additivity.

EXAMPLES: 4) *Banach's match boxes.* In an address in honour of Banach, Hugo Steinhaus made humorous reference to the smoking habits of the famous mathematician. He may not have anticipated that his off-hand reference would become part of the established lore of the field.

An eminent mathematician fuels a smoking habit by keeping matches in both trouser pockets. When impelled by need he reaches a hand into a randomly selected pocket and grubs about for a match. Suppose he starts with n matches in each pocket. What is the probability that when he first discovers a pocket to be empty of matches the other pocket contains exactly k matches? Suppose he discovers that the right pocket is empty. Then he has made a total of $(n + 1) + (n - k)$ forays into the two pockets culminating in the $(n + 1)$ th abortive visit to the right pocket. Model this as a sequence of Bernoulli trials with success probability $p = 1/2$ of visiting the right pocket and failure probability $q = 1/2$ of visiting the left pocket. Then the probability in question is the probability that the waiting time to the $(n + 1)$ th success is $n - k$ and is hence given by the negative binomial

$$w_{n+1}(n - k; 1/2) = \binom{2n - k}{n - k} 2^{-2n+k-1} = \binom{2n - k}{n} 2^{-2n+k-1}.$$

Of course, an identical situation prevails if he happens to discover the left pocket empty. It follows that at the moment of truth when he first discovers an empty pocket, the probability that the other pocket contains k matches is $\binom{2n - k}{n} 2^{-2n+k}$.

5) *The problem of the points.* If a game of chance where the gamblers put up stakes is abandoned before conclusion, how should the pot be split? This is the famous *problem of the points* which was posed to Blaise Pascal by the Chevalier de Méré in 1654. The problem was satisfactorily resolved in an historic correspondence between Pascal and Pierre Fermat. They proposed the eminently reasonable idea that the amount credited to each gambler should be based not on past history but proportional to his chances of winning at the moment the game was abandoned. In other words, the past history of the game is not important; what matters is the range of possible future directions the game might have taken were it not interrupted. This exchange of letters between the two luminaries is frequently credited with sparking interest and development in the nascent science. Here is the model problem.

A game of chance between two individuals is decided based on the results of a series of independent trials. The players accumulate points by winning trials. At the time the game is interrupted, our gamine protagonist requires m points to win the game, her opponent n . If our gambler wins any given trial with probability p and wins and losses are independent across trials, what is the probability $P_{m,n}(p)$ she would have won the game if it were played to a conclusion?

Let us consider a sequence of Bernoulli trials with success probability p , each trial representing the outcome of a game. The gambler wins the contest if, when the m th success occurs, there are fewer than n failures up to that point. In other words, she wins the contest if the number of failures before the m th success satisfies $0 \leq k \leq n - 1$. As cast it is obvious that this is a waiting time problem with probability

$$P_{m,n}(p) = \sum_{k=0}^{n-1} w_m(k; p) = \sum_{k=0}^{n-1} \binom{-m}{k} (-q)^k p^m. \quad (7.4)$$

If the two players start on even terms and the trials are fair, that is $p = q = 1/2$, we obtain $P_{n,n}(1/2) = 1/2$ as is intuitive and just. Pascal's and Fermat's approaches are sketched in the *Problems* at the end of this chapter. ►

The variable T_r has a simple and direct interpretation. We again consider an uninterrupted run of Bernoulli trials with success probability p and interpret W_1 as the number of failures before the first success, W_2 as the number of failures between the first and second successes, and so on, with W_r being the number of failures between the $(r-1)$ th and r th successes. It is clear then that T_r may be identified with the total number of failures before the r th success or, more evocatively, the *waiting time till the r th success*. Its distribution may now be given a direct combinatorial interpretation by observing that $T_r = k$ if, and only if, the $(r+k)$ th trial results in success and the first $r+k-1$ trials contain exactly $r-1$ successes (or, equivalently, k failures). The probability of the former eventuality is p , that of the latter $\binom{r+k-1}{r-1} p^{r-1} q^k$; as the trials are independent we may multiply the two probabilities to obtain (7.3).



8 Run lengths, quality of dyadic approximation

Suppose X_1, X_2, \dots is a sequence of symmetric Bernoulli trials. For each n , let R_n represent the length of the run of failures beginning at trial n : the event $R_n = r$ occurs if, and only if, $X_n = X_{n+1} = \dots = X_{n+r-1} = 0$ and $X_{n+r} = 1$. Thus, R_n is governed by the geometric distribution $P\{R_n = r\} = 2^{-r-1}$ or, equivalently, $P\{R_n \geq r\} = 2^{-r}$ for $r \geq 0$. A little introspection shows that the maximal run length over a given number of trials n cannot increase without bound as n increases. Suppose $\{r_n, n \geq 1\}$ is any positive sequence. Then, by the first Borel–Cantelli lemma (Theorem IV.4.2), $P\{R_n \geq r_n \text{ i.o.}\} = 0$ if $\sum_n 2^{-r_n}$ converges. The divergence of the harmonic series suggests now that the critical run length sequence should be around $\log_2 n$. Indeed, fix any $\epsilon > 0$ and set $r_n = \lceil (1 + \epsilon) \log_2 n \rceil$. Then $\sum_n 2^{-r_n} \leq \sum_n n^{-1-\epsilon}$ converges. As $\epsilon > 0$ may be arbitrarily specified, we obtain

$$P\left\{\limsup_n \frac{R_n}{\log_2 n} > 1\right\} = 0, \text{ or, equivalently, } P\left\{\limsup_n \frac{R_n}{\log_2 n} \leq 1\right\} = 1. \quad (8.1)$$

Thus, the maximal run length sequence cannot exceed $(1 + \epsilon) \log_2 n$ more than a finite number of times. This establishes an *upper envelope* of $\log_2 n$ for the run length sequence.

On the other hand, it is clear that arbitrarily large runs must sooner or later occur in an infinite number of trials. What, for instance, is the chance that a run of length r eventually occurs? We cannot use the second Borel–Cantelli lemma (Theorem IV.4.3) directly because the events $\{R_n \geq r\}$ are not independent. But as the event $\{R_n \geq r\} = \{X_n = X_{n+1} = \dots = X_{n+r-1} = 0\}$ is determined completely by the block of r variables $X_n, X_{n+1}, \dots, X_{n+r-1}$, the events $A_k = \{R_{k+r-1} \geq r\}$ for $k = 0, 1, \dots$ are determined by non-overlapping, finite blocks of Bernoulli trials and are hence independent. Each A_k has probability 2^{-r} and so the sum $\sum_k P(A_k) = \sum_k 2^{-r}$ is divergent. By the second Borel–Cantelli lemma it follows that $P\{A_k \text{ i.o.}\} = 1$ and, *a fortiori*, $P\{R_n \geq r \text{ i.o.}\} = 1$. As r may be selected arbitrarily large, this shows that large runs must occur. By a slight tweak of this argument we can now hone in on the critical rate of increase of the maximal run lengths.

Suppose $\{r_n, n \geq 1\}$ is an increasing sequence of positive integers. We may partition the sequence of Bernoulli trials into successive blocks of r_n non-overlapping trials by the following recursive procedure. Set $n_1 = 1$ and, for $k \geq 1$, establish the recurrence $n_{k+1} = n_k + r_{n_k}$. Then the occurrence of the event $\{R_{n_k} \geq r_{n_k}\}$ is determined completely by the r_{n_k} trials $X_{n_k}, X_{n_k+1}, \dots, X_{n_{k+1}-1}$ in the k th block. As these blocks of trials are non-overlapping, the events $A_k = \{R_{n_k} \geq r_{n_k}\}$ for $k = 1, 2, \dots$ are independent. The stage is set for the second Borel–Cantelli lemma. We have

$$\begin{aligned} \sum_{k=1}^{\infty} P\{R_{n_k} \geq r_{n_k}\} &= \sum_{k=1}^{\infty} 2^{-r_{n_k}} = \sum_{k=1}^{\infty} 2^{-r_{n_k}} \left(\frac{n_{k+1} - n_k}{r_{n_k}} \right) \\ &= \sum_{k=1}^{\infty} \sum_{n_k \leq n < n_{k+1}} \frac{2^{-r_{n_k}}}{r_{n_k}} \stackrel{(*)}{\geq} \sum_{k=1}^{\infty} \sum_{n_k \leq n < n_{k+1}} \frac{2^{-r_n}}{r_n} = \sum_{n=1}^{\infty} \frac{2^{-r_n}}{r_n}. \end{aligned}$$

The step marked $(*)$ follows because the sequence $\{r_n\}$ is increasing and, in particular, $r_{n_k} \leq r_n$ for $n_k \leq n < n_{k+1}$. Specialising by setting $r_n = \lfloor \log_2 n \rfloor$, we see that $\sum_n 2^{-r_n}/r_n \geq \sum_n 1/(n \log_2 n)$ and the series diverges. By the second Borel–Cantelli lemma it follows that $P\{R_{n_k} \geq \log_2 n_k \text{ i.o.}\} = 1$, *a fortiori*, $P\{R_n \geq \log_2 n \text{ i.o.}\} = 1$, and so

$$P\left\{ \limsup_n \frac{R_n}{\log_2 n} \geq 1 \right\} = 1. \quad (8.2)$$

We hence have a *lower envelope* of $\log_2 n$ for the run length sequence. Putting (8.1) and (8.2) together we obtain a compact and beautiful result.

THEOREM 1 *The run length sequence $\{R_n, n \geq 1\}$ satisfies*

$$P\left\{ \limsup_n \frac{R_n}{\log_2 n} = 1 \right\} = 1.$$

In words, the maximal run length has asymptotic order *precisely* $\log_2 n$.

In view of the intimate connection between coin tosses and the digits of a dyadic expansion that was fully fleshed out in Chapter V, results on run lengths imply corresponding results for the quality of approximation of dyadic expansions. Suppose $0 \leq t < 1$ is a point in the unit interval. We begin with the dyadic expansion (V.2.1), repeated here for convenience. Write $t = \sum_{k=1}^{\infty} z_k 2^{-k}$ where $z_k \in \{0, 1\}$ and, by convention, we again use the terminating expansion when t has two representations. We may

hence put t in one-to-one correspondence with a binary sequence $(z_n, n \geq 1)$ and, echoing our earlier notation, we write $R_n(t)$ for the run length of zeros starting at location n in the binary representation.

Now, for each n , let $t_n = \sum_{k=1}^n z_k 2^{-k}$ be the dyadic expansion of t truncated to n terms and write $\delta_n(t) = t - t_n$ for the *dyadic truncation error to n terms*. It is clear that $0 \leq \delta_n(t) \leq 2^{-n}$ but more can be said. Introduce the nonce notation $\rho_n(t) = \delta_n(t)/2^{-n}$ for the truncation error relative to the maximum. We may then write

$$\rho_n(t) = \frac{1}{2^{-n}} \sum_{k=n+1}^{\infty} z_k 2^{-k} = \sum_{j=1}^{\infty} z_{j+n} 2^{-j}.$$

The expression on the right is the dyadic expansion of the relative error $\rho_n(t)$ and it is clear that it begins with $R_{n+1}(t)$ zeros followed by a one. Consequently,

$$2^{-R_{n+1}(t)-1} \leq \rho_n(t) \leq 2^{-R_{n+1}(t)}.$$

Suppose to begin that $\rho_n(t) \leq x_n$ where $\sum_n x_n$ converges; as a specific example we may select $x_n = n^{-1-\epsilon}$ for any fixed $\epsilon > 0$. Then the lower bound on the relative error shows that $\sum_n 2^{-R_n(t)}$ converges. Writing λ as usual for the Lebesgue measure (length) of subsets of the unit interval, by the correspondence between binary digits and coin tosses it follows that $\lambda\{t : \rho_n(t) \leq n^{-1-\epsilon} \text{ i.o.}\} = 0$. On the other hand, if $R_{n+1}(t) \geq \log_2 n$ then $\sum_n 2^{-R_n(t)}/R_n(t)$ diverges. By the argument leading up to (8.2) it follows that $\lambda\{t : \rho_n(t) \leq n^{-1} \text{ i.o.}\} = 1$.

THEOREM 2 *Let $\delta_n(t)$ denote the dyadic truncation error to n terms of any point t in the unit interval $[0, 1]$. Then, for every $\epsilon > 0$, we have*

$$\lambda\{t : \delta_n(t) \leq \frac{2^{-n}}{n^{1+\epsilon}} \text{ i.o.}\} = 0 \text{ and } \lambda\{t : \delta_n(t) \leq \frac{2^{-n}}{n} \text{ i.o.}\} = 1.$$

Thus, for almost every point in the unit interval, the dyadic truncation error falls below $2^{-n}/n$ infinitely often but below $2^{-n}/n^{1+\epsilon}$ only a finite number of times.

9 The curious case of the tennis rankings

The player rankings maintained by the Association of Tennis Players (ATP) show that a small handful of professionals maintain a seeming stranglehold on top of the world rankings for long periods of time. But even the casual watcher of games is aware that in the top flight points are fiercely competitive and the point-by-point difference between an elite player and one in the next rung is minute. If a top player is even slightly off-colour his opponent is poised to take advantage and one might anticipate a frequent shuffling in the world order. What then might explain the relatively unchanging rankings? While several factors may be proposed in explanation of this phenomenon—the weighted average of past performance that is used by the ATP, for instance, assuredly plays a part—let us examine whether the *structure* of the game itself may be a contributory factor.

Tennis, like many other games of its ilk, is played in a succession of points. Consider for instance a match between two players, say \mathfrak{A} and \mathfrak{B} . If we write 1 for a point won by \mathfrak{A} and 0 for a point won by \mathfrak{B} , the progression of points may then be represented by a sequence such as

100111011111001111110100001001101101000101001101001100111

at the conclusion of which \mathfrak{A} was declared the winner. A count of points won shows that in this (abbreviated) match \mathfrak{A} won 34 out of 59 points played. A seductively simple analysis runs as follows. If one were to imagine that an (interminable) succession of matches played by our intrepid protagonists were stacked back-to-back then one would obtain a sequence of points with a point every so often identified as concluding a match and initiating the next. Suppose that \mathfrak{A} wins 55% of the points played. If we focus on the outcome of the 59th point in the sequence, then the chances of \mathfrak{A} winning it are 55%, those of \mathfrak{B} winning it are 45%, and, as a match was decided on this point these odds should also mirror whether \mathfrak{A} or \mathfrak{B} won the match. From this perspective \mathfrak{B} should win her fair share of the matches.

The analysis fails to take into account the special structure of the contest. A progression of points is divided into a succession of *games*, *sets*, and, finally, *matches*: points determine games, games determine sets, and sets determine matches.

To begin, each game is initiated with the two players each reinitialising their point totals to zero. At the conclusion of each point during the progress of the game, the player who wins the point increments his point total by one. A player wins the game at the *first instant* when she leads the other player by at least two points *and* has won at least four points.¹⁰ If we represent point totals for \mathfrak{A} first, \mathfrak{B} second, game-winning totals for player \mathfrak{A} are of the form $(4, 0)$, $(4, 1)$, $(4, 2)$, $(5, 3)$, $(6, 4)$, etc., with tallies reversed for wins by \mathfrak{B} . Once a game has been decided in favour of one or the other player, the process is now begun anew for the next game. For example, during the progression of the first game in the sample sequence of points given above the point totals through the first six points are $(1, 0)$, $(1, 1)$, $(1, 2)$, $(2, 2)$, $(3, 2)$, $(4, 2)$, at which point \mathfrak{A} wins the game. A longer typical progression of points occurs at the end of the given sequence and tallies $(1, 0)$, $(1, 1)$, $(1, 2)$, $(2, 2)$, $(3, 2)$, $(3, 3)$, $(4, 3)$, $(4, 4)$, $(4, 5)$, $(5, 5)$, $(6, 5)$, $(6, 6)$, $(6, 7)$, $(7, 7)$, $(8, 7)$, $(9, 7)$ with the final burst of three points securing the game for \mathfrak{A} at the sixteenth and final point of the game.

Game totals are initially set equal to zero and incremented by one for each player each time she wins a game. A set is won at the *first instant* one

¹⁰Standard game terminology calls 0 points “love” and the first three points “fifteen”, “thirty”, and “forty”, respectively. Ties at “forty” and subsequently are called “deuce” with one player or the other getting “advantage” at the conclusion of the next point. If the player with “advantage” in such a situation wins the subsequent point, she also wins the “game”. The situation as described is identical to our streamlined point system and we will not worry about the peculiar conventions of the real game.

player leads the other by at least two games *and* has accumulated at least six games during the progress of the set. Thus the progression of games determines sets.

Finally, set totals are incremented and game totals reset to zero each time a player wins a set. The match is won at the *first instant* one player wins three sets. Once a match is decided, set and game counts are reinitialised to zero and the process started anew for the next match.

One may imagine a succession of points flagged at intervals by *game markers* (say, \uparrow or \downarrow in favour of \mathfrak{A} and \mathfrak{B} , respectively) which indicate the points at which a game was concluded, the subsequence of points determining games flagged at intervals by *set markers* (say, \Uparrow or \Downarrow in favour of \mathfrak{A} and \mathfrak{B} , respectively) that indicate the game points where sets were concluded, and, finally, the subsequence of points determining sets flagged at intervals by *match markers* (say, \circlearrowleft or \circlearrowright in favour of \mathfrak{A} and \mathfrak{B} , respectively) that indicate the set points where matches were concluded. The progression of points given earlier with game and set markers inserted into the sequence yields

$$100111\uparrow^{(1,0)}01111\uparrow^{(2,0)}100111\uparrow^{(3,0)}1111\uparrow^{(4,0)}01000\downarrow^{(4,1)} \\ 010011011011\uparrow^{(5,1)}00010\downarrow^{(5,2)}1001101001100111\uparrow^{(6,2),(1,0)}$$

at the conclusion of which \mathfrak{A} has won the first set of the match with a game total of $(6, 2)$.

Let us strip the game to its essence by supposing that \mathfrak{A} has probability p of winning any given point and that the outcome of each point in the game represents a statistically independent experiment. The succession of points hence forms a sequence of Bernoulli trials with fixed parameter p . The knowledgeable player-reader knows, of course, that this represents a gross simplification: the (alternating) server usually maintains a decided edge for the duration of her service game; fatigue, physical and mental, as the match progresses may lead to changes in point odds; the supposed ability of an elite player to raise their game on the “big” points may alter odds. I will leave it to the reader to make up her own mind how taking any of these factors into account may change the analysis.

With points forming a sequence of Bernoulli trials with success probability p , let p_r denote the probability that \mathfrak{A} wins a game with a point total of r . It will be convenient to introduce the notation (m, n) to represent the event that the game terminates after completion of the $(m + n)$ th point with the accumulated point score (m, n) . For $r = 4$, as \mathfrak{A} wins with a point total of four if she gets to four points before her opponent gets to three, we recognise an incarnation of the problem of the points, whence from (7.4) we obtain $p_4 = P_{4,3}(p) = p^4 + 4qp^4 + 10q^2p^4$. For $r \geq 5$ the occurrence of the event $(r, r - 2)$ is predicated upon ties at $(3, 3), (4, 4), \dots, (r - 2, r - 2)$, followed by

two successive point wins by \mathfrak{A} . It follows that, for each $r \geq 5$,

$$p_r = P\{\overline{(r, r-2)}\} = \left\{ \binom{6}{3} q^3 p^3 \right\} \left\{ \binom{2}{1} qp \right\}^{r-5} p^2 = 20q^3 p^5 (2qp)^{r-5}.$$

The probability $g(p)$ that \mathfrak{A} wins the game is hence given by additivity of probability measure to be

$$\begin{aligned} g(p) &= \sum_{r=4}^{\infty} p_r = p^4 + 4qp^4 + 10q^2p^4 + 20q^3p^5 \sum_{r=5}^{\infty} (2qp)^{r-5} \\ &= p^4 \left(1 + 4q + 10q^2 + \frac{20q^3p}{1-2qp} \right) = \frac{p^4(15 - 34p + 28p^2 - 8p^3)}{1 - 2p + 2p^2}. \end{aligned}$$

For $p = 1/2$ the expression evaluates to $1/2$, as it must.

Thus, in our model, games are won by \mathfrak{A} with probability $g(p)$ and the succession of games which determines a set now forms a sequence of Bernoulli trials with success probability $g = g(p)$. It is clear that the instant a set is won is another waiting time problem, the analysis mirroring that for a single game. In an abuse of notation, write $s = s(g) = s(g(p))$ for the probability that \mathfrak{A} wins a set. We then obtain

$$\begin{aligned} s(g) &= P_{6,5}(g) + \binom{10}{5} (1-g)^5 g^7 \sum_{k \geq 0} [2(1-g)]^k \\ &= \frac{g^6(210 - 888g + 1545g^2 - 1370g^3 + 616g^4 - 112g^5)}{1 - 2g + 2g^2}. \end{aligned}$$

The sets themselves form a sequence of Bernoulli trials with success probability $s = s(g) = s(g(p))$. A match never needs more than five sets to determine the winner, the instant a match is won forming yet another waiting time problem. Again, the analysis mirrors that for a single game. Abusing notation, write $m = m(s) = m(s(g(p)))$ for the probability that \mathfrak{A} wins the match. Then

$$m(s) = P_{3,3}(s) = s^3 + 3(1-s)s^3 + 6(1-s)^2s^3 = s^3(10 - 15s + 6s^2).$$

The resolution of the rankings paradox thus depends, at least in part, upon the game structure. The point at which a match is decided is not fixed *a priori* but is decided by the unfolding sequence of points themselves. That is to say, the decision point is not typical of points that are played, as naïve intuition might have suggested, but is special, identified by the progression of past points in accordance with the particular rules of the game. The probabilities of winning a game, set, and match are shown graphed as a function of the point-winning probability p in Figure 5. The highly non-linear character of these functions is noteworthy. In particular, $m \neq p$ (as one might perhaps have naïvely expected) except at the special points $p = 0$, $p = 1/2$, and $p = 1$. Even more remarkable is the threshold phenomenon that one sees emerging around $p = 1/2$ when one moves from games to sets to matches. A small edge p slightly

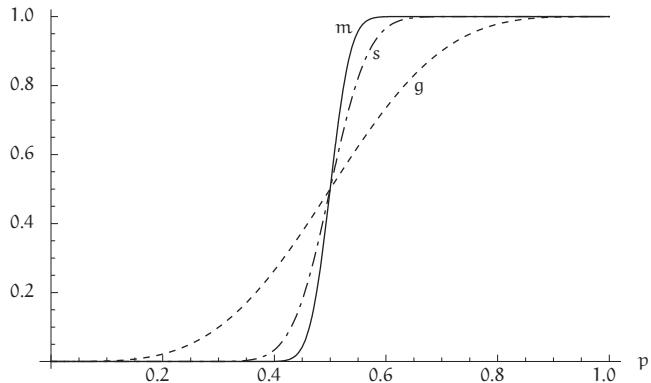


Figure 5: Game, set, and match win probabilities in tennis.

in excess of $1/2$ in point win probabilities is *boosted* to a significantly larger edge in games $g = g(p)$ which, in turn, is boosted to a much larger edge in sets $s = s(g(p))$ which, finally, is boosted to a very pronounced edge in matches $m = m(s(g(p)))$. Thus, a slight edge of 55% in points sees a dramatic boosting of the advantage to 62% in games, 82% in sets, and an astounding 96% in match win percentages!

In our admittedly simplified model, a player need maintain only a very small edge on a point-by-point basis to build up a very large advantage in match wins. This might explain, in part, why we see a very small handful of players atop the world rankings on a year-in, year-out basis. While the separation between professionals on a point-by-point basis may be very modest, the best players may maintain a very small, but distinct statistical edge over the others on each point. And this small edge is boosted by the game's structure into a sustained match-winning edge.

10 Population size, the hypergeometric distribution

The hypergeometric distribution appears in the context of sampling from mixed populations. In its simplest form, consider a population of n elements comprised of two subpopulations of sizes m and $n - m$, respectively. A random sample of r elements from the population will contain exactly i elements from the first subpopulation (and, of course, $r - i$ elements from the second subpopulation) with probability

$$h(i) = h_{m;n}(i; r) = \frac{\binom{m}{i} \binom{n-m}{r-i}}{\binom{n}{r}} \quad (0 \leq i \leq r). \quad (10.1)$$

Vandermonde's convolution, $\sum_i \binom{m}{i} \binom{n-m}{r-i} = \binom{n}{r}$ [see Problem 1], shows that this is indeed a distribution.

EXAMPLES: 1) *Quality control.* A production lot of n widgets contains an unknown number m of defective widgets. A quality inspection selects r widgets at random for examination and finds that i are defective. How can one proceed to estimate the value of m that is most consistent with this data?

For fixed n and r , the probability that i defective widgets are found in a random sample of size r depends only on the unknown number of defective widgets in the population via the hypergeometric distribution (10.1). With n , r , and i fixed, in a slight abuse of notation write $h_m = \binom{m}{i} \binom{n-m}{r-i} / \binom{n}{r}$ for this probability to emphasise the dependence on the unknown m . Following Fisher's principle of maximum likelihood, it is natural to estimate the unknown number of defective widgets by $\hat{m} = \arg \max_m h_m$, the value of m that yields the largest probability of observations consistent with the data. In our example, Fisher's principle yields a simple and intuitive answer. As

$$\frac{h_m}{h_{m-1}} = \frac{m(n-m-r+i+1)}{(m-i)(n-m+1)},$$

simple manipulations show that $h_m > h_{m-1}$ if $m < i(n+1)/r$ and $h_m < h_{m-1}$ if $m > i(n+1)/r$. Thus, h_m takes its maximum value when m is the largest integer short of $\frac{i}{r}(n+1)$. If a lot of 5000 widgets is produced, 100 examined, and 10 found defective, the maximum likelihood estimate of the total number of defective widgets in the sample is 500 or 10% of the whole. The distribution (10.1) provides the statistician more detailed information with which to test the reasonableness of the estimate.

2) *Population size estimates from recapture data.* Zoological censuses encounter particular difficulties because of the pronounced disinclination of the members of the population (birds, fish, etc.) to participate in the census. A widely used approach to estimate unknown animal populations utilises recapture data. A number of animals, say m , are captured, marked, and released. A second random sample of r animals are then caught subsequently and an examination of the markings show that i of these are specimens that have been recaptured. What can be said about the unknown number of animals in the total population?

With natural assumptions on the independence of the samples, given m , r , and i , the probability that i members of the original sample of size m are found to have been recaptured in the new sample of size r depends only on the unknown population size n and is given again by the distribution (10.1). In another abuse of notation, write $h_n = \binom{m}{i} \binom{n-m}{r-i} / \binom{n}{r}$ for this probability to emphasise the dependence now on n . Fisher's maximum likelihood principle suggests then that we estimate the unknown population by $\hat{n} = \arg \max_n h_n$ as

the population size that best fits the observed data. The maximum likelihood principle again takes a simple and intuitive form in this setting. It is easy to verify that

$$\frac{h_n}{h_{n-1}} = \frac{(n-m)(n-r)}{n(n-m-r+i)}$$

and the right-hand side is greater than 1 if, and only if, $n < \frac{r}{i}m$. Thus, h_n achieves its maximum value at the largest integer short of $m/\frac{i}{r}$. If a second sample of size $r = 100$ is found to contain $i = 10$ recaptures then the maximum likelihood estimate places the animal population as ten times larger than the size of the first sample. Again, the specific form of the distribution (10.1) permits statistical testing of the conclusion. ►

With the parameters m , n , and r held fixed, the ratio of successive terms of the hypergeometric distribution is given by

$$\frac{h(i)}{h(i-1)} = \frac{(m-i+1)(r-i+1)}{i(n-m-r+i)}$$

and simple manipulations show then that $h(i) > h(i-1)$ if $i < (m+1)(r+1)/(n+2)$ and $h(i) < h(i-1)$ if $i > (m+1)(r+1)/(n+2)$. It follows hence that $h(i)$ achieves its maximum value at the largest integer short of $\frac{m+1}{n+2}(r+1)$. Thus, for large populations, in a random sample of size r , it is most likely that the two subpopulations are represented approximately in the ratios $\frac{m}{n}r$ and $(1 - \frac{m}{n})r$, respectively, as is reasonable and just.

The general hypergeometric distribution obtains for sampling without replacement from a population of size n consisting of k distinct subpopulations of sizes n_1, \dots, n_k . (Of course, $n_1 + \dots + n_k = n$.) In a random sample of size r from the population, let X_1, \dots, X_k , respectively, denote the number of elements from each of the k subpopulations. The k -tuple (X_1, \dots, X_k) then has the hypergeometric distribution

$$P\{X_1 = i_1, \dots, X_k = i_k\} = \frac{\binom{n_1}{i_1} \cdots \binom{n_k}{i_k}}{\binom{n}{r}}$$

where i_1, \dots, i_k are positive integers adding up to r .

EXAMPLE 3) Bridge. A bridge hand consists of 13 cards from a 52-card deck consisting of the four suits ♠, ♥, ♦, and ♣, each comprised of 13 cards. The number of cards in the hand in each of the four suits is a four-tuple $(X_{\spadesuit}, X_{\heartsuit}, X_{\diamondsuit}, X_{\clubsuit})$ with distribution $\binom{13}{i_1} \binom{13}{i_2} \binom{13}{i_3} \binom{13}{i_4} / \binom{52}{13}$. The probability that one suit is void is

$$4 \sum_{i_1+i_2+i_3=13} \binom{13}{i_1} \binom{13}{i_2} \binom{13}{i_3} / \binom{52}{13} = 4 \binom{39}{13} / \binom{52}{13} = .051163 \dots$$

as the sum on the left may be identified with the number of ways 13 cards may be drawn from the 39 cards that remain after one suit is eliminated from consideration. The probability that two suits are void, likewise, is given by

$$\binom{4}{2} \sum_{i_1+i_2=13} \binom{13}{i_1} \binom{13}{i_2} / \binom{52}{13} = 6 \binom{26}{13} / \binom{52}{13} = .000098 \dots$$

A void in one suit may be expected 5% of the time. Two voids are extremely unlikely. ►

11 Problems

1. *Vandermonde's convolution.* For any positive integers μ , ν , and n , prove by induction that

$$\binom{\mu}{0} \binom{\nu}{n} + \binom{\mu}{1} \binom{\nu}{n-1} + \dots + \binom{\mu}{n} \binom{\nu}{0} = \binom{\mu+\nu}{n}.$$

[First prove the result for $\mu = 1$ and all ν .] Then provide a direct proof by a probabilistic argument. Alexandre Vandermonde wrote about this identity in 1792.

2. *Continuation.* Show that, for every positive integer n ,

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \binom{n}{2}^2 + \dots + \binom{n}{n}^2 = \binom{2n}{n}.$$

3. *Closure of the binomial under convolutions.* Show directly from the binomial sums that $(b_\mu * b_\nu)(k) = b_{\mu+\nu}(k)$ where, for each n , b_n represents a binomial distribution corresponding to n tosses of a coin with fixed success probability p . Now provide a direct probabilistic argument.

4. Suppose X_1 and X_2 are independent Poisson variables of means λ_1 and λ_2 , respectively. Determine the conditional distribution of X_1 given that $X_1 + X_2 = n$.

5. *The multinomial distribution.* Suppose $\{p_1, \dots, p_r\}$ forms a discrete probability distribution. Write

$$p(k_1, k_2, \dots, k_r) = \frac{n!}{k_1! k_2! \dots k_r!} p_1^{k_1} p_2^{k_2} \dots p_r^{k_r}. \quad (11.1)$$

By summing over all positive integers k_1, k_2, \dots, k_r satisfying $k_1 + k_2 + \dots + k_r = n$, show that $\sum p(k_1, k_2, \dots, k_r) = 1$ and hence that the values $p(k_1, k_2, \dots, k_r)$ determine a discrete probability distribution on a lattice in r dimensions. This is the multinomial distribution; for $r = 2$ we recover the binomial distribution.

6. *Central term of the multinomial.* Show that the maximal term of the multinomial distribution (11.1) satisfies the inequalities

$$np_j - 1 < k_j \leq (n+r-1)p_j \quad (1 \leq j \leq r).$$

[Hint: Show that a maximal term of (11.1) satisfies $p_i k_j \leq p_j (k_i + 1)$ for every pair (i, j) . Obtain the lower bound by summing over all j , the upper bound by summing over all $i \neq j$. W. Feller attributes the result and the neat proof to P. A. P. Moran.]

7. *Central term of the trinomial.* Show that the trinomial $\frac{(3n)!}{j!k!(3n-j-k)!} 3^{-3n}$ achieves its maximum value when $j = k = n$.
8. In a variant of Pepys's problem, which is more likely: that when twelve dice are tossed each of the face values 1 through 6 show exactly twice or that when six dice are tossed each of the face values show exactly once?
9. *Ones and sixes.* Let X and Y denote the number of 1s and 6s, respectively, that turn up in n independent throws of a fair die. What is the expected value of the product XY ? [Problem VII.5 simplifies calculations.]
10. *Traffic.* A pedestrian can cross a street at epochs $k = 0, 1, 2, \dots$. The event that a car will be passing the crossing at any given epoch is described by a Bernoulli trial with success probability p . The pedestrian can cross the street only if there will be no car passing over the next three epochs. Find the probability that the pedestrian has to wait exactly $k = 0, 1, 2, 3, 4$ epochs.
11. *Log convexity.* We say that a function p defined on the integers is *log convex* if it satisfies $p(k-1)p(k+1) \leq p(k)^2$ for all k . Show that the binomial distribution $b_n(k; p)$ and the Poisson distribution $p(k; \lambda)$ are both log convex.
12. *Continuation.* Find a distribution $p(k)$ on the positive integers for which equality holds in the log convex inequality, i.e., $p(k-1)p(k+1) = p(k)^2$ for all $k \geq 1$.
13. A book of 1000 pages contains 1000 misprints. Estimate the chances that a given page contains at least three misprints.
14. *Poker.* A *royal flush* in poker is a hand comprised of the cards ace, king, queen, jack, and ten, all from the same suit. Determine its probability. How large does n have to be so that the chances of at least one royal flush in n hands are at least $1 - e^{-1} \approx 2/3$?
15. Show that the terms $p(k; \lambda) = e^{-\lambda} \lambda^k / k!$ of the Poisson distribution attain their maximum value when $k = \lfloor \lambda \rfloor$.
16. Show that the sequence $a_k = b_n(k; p) / p(k; np)$ attains its maximum value for $k = \lfloor \lambda + 1 \rfloor$.
17. *Differentiating under the summation sign.* Write $S(q) = \sum_{k=0}^{\infty} q^k$ for $0 < q < 1$. Show that it is permissible to differentiate under the summation sign to form a series representation for $S'(q)$. Hence, by differentiating both sides of the geometric series formula obtain the mean of the geometric distribution $w(k; p)$.
18. *Continuation.* By differentiating both sides of the geometric series formula twice derive the variance of $w(k; p)$.
19. Suppose X_1 and X_2 are independent with the common geometric distribution $w(k; p)$. Determine the conditional distribution of X_1 given that $X_1 + X_2 = n$.
20. *Waiting for Mr. Right.* A game of chance proceeds by repeated throws of an n -sided die (with face values 1, ..., n) resulting in a sequence of values X_1, X_2, \dots . A gambler pays a house fee f to enter the game and thereafter an additional fee a for every trial that she stays in the game. She may leave the game at any time and at departure receives as payoff the last face value seen before she quits. Thus, if she leaves after τ trials she will have paid $f + a\tau$ in accumulated fees and receives a payoff X_τ ; her winnings (negative winnings are losses) are $W_\tau = X_\tau - (f + a\tau)$. The gambler who plays a predetermined number of times has expected winnings $(n+1)/2 - (f + a\tau)$.

decreasing monotonically with τ so that there is no incentive to play for more than one trial. The reader who has done the marriage problem II.28, 29, however, may feel that our intrepid gambler can benefit from a properly selected waiting time strategy. Fix any integer $0 \leq r \leq n - 1$ and let τ_r be the first trial for which $X_{\tau_r} > r$. Determine the expected winnings $E(W_{\tau_r})$ of the gambler who plays till the first instant she sees a value $r + 1$ or higher and then quits, and optimise over r to maximise her expected winnings. Suppose n is large. Should she play the game if the house charges an entrance fee $f = 9n/10$ and a playing fee $a = 1$ for each trial? In this case, what would a fair entrance fee f be? If $f = 0$ what would a fair playing fee a be?

21. *Negative binomial.* Determine the mean and the variance of the negative binomial distribution $w_r(\cdot; p)$ by direct computation. Hence infer anew the mean and variance of the geometric distribution. [Hint: The representation $\binom{-r}{k}(-q)^k p^r$ for the negative binomial probabilities makes the computation simple.]
22. In Banach's match box problem (Example 7.4) determine the probability that at the moment the first pocket is emptied of matches (as opposed to the moment when the mathematician reaches in to find an empty pocket) the other contains exactly r matches.
23. A sequence of Bernoulli trials with success probability p is continued until the r th success is obtained. Suppose X is the number of trials required. Evaluate $E(r/X)$. (The infinite series that is obtained can be summed in closed form.)
24. *The problem of the points.* In a succession of Bernoulli trials with success probability p , let $P_{m,n}$ denote the probability that m successes occur before n failures. Show that $P_{m,n} = pP_{m-1,n} + qP_{m,n-1}$ for $m, n \geq 1$ and solve the recurrence. This solution is due to Blaise Pascal.
25. *Continuation.* Argue that for m successes to occur before n failures it is necessary and sufficient that there be at least m successes in the first $m + n - 1$ trials and thence write down an expression for $P_{m,n}$. This solution is due to Pierre Fermat.
26. *Equally matched tennis players.* Suppose players \mathfrak{A} and \mathfrak{B} alternate serve from point to point in a modified tie-break in a tennis match, $p_{\mathfrak{A}}$ and $p_{\mathfrak{B}}$, respectively, the probabilities of their winning the point on their serve. The tie-break is won by whichever player first takes a two-point lead. If $p_{\mathfrak{A}} = 0.99$ and $p_{\mathfrak{B}} = 0.96$ then the game will go on for a very long time with players holding serve. The first person to lose her serve will, with high probability, lose the game. Show that the probability that \mathfrak{A} wins is approximately $4/5$. (This pretty observation is due to T. M. Cover.)
27. *Alternating servers in tennis.* In singles play in tennis the player serving alternates from game to game. Model the fact that the server usually has a slight edge on points during her serve by setting \mathfrak{A} 's chances of winning a point as p_1 on her serve and p_2 on her opponent's serve. Determine the game, set, and match win probabilities for \mathfrak{A} under this setting.
28. *An approximation for the hypergeometric distribution.* A large population of N elements is made up of two subpopulations, say, red and black, in the proportion $p : q$ (where $p + q = 1$). A random sample of size n is taken without replacement. Show that, as $N \rightarrow \infty$, the probability that the random sample contains exactly k red elements tends to the binomial probability $b_n(k; p)$.

29. *Capture-recapture.* A subset of m animals out of a population of n has been captured, tagged, and released. Let R be the number of animals it is necessary to recapture (without re-release) in order to obtain k tagged animals. Show that R has distribution

$$P\{R = r\} = \frac{m}{n} \binom{m-1}{k-1} \binom{n-m}{r-k} / \binom{n-1}{r-1}$$

and determine its expectation.

The following problems deal with random walks.

30. *Enumerating paths.* If $a > 0$ and $b > 0$, the number of sample paths of a random walk satisfying $S_1 > -b, \dots, S_{n-1} > -b$, and $S_n = a$ equals $N_n(a) - N_n(a + 2b)$. If $b > a$, how many paths are there whose first $n - 1$ steps lie strictly below b and whose n th step culminates in a ?

31. *Regular parenthetical expressions.* The use of parentheses in mathematics to group expressions and establish the order of operations is well understood. The rules governing what constitutes a proper collection of left and right parentheses in an expression are simple: (i) the number of left parentheses must equal the number of right parentheses; and (ii) as we go from left to right, the number of right parentheses encountered can never exceed the number of left parentheses previously encountered. We say that an expression consisting of left and right parentheses is regular (or well-formed) if it satisfies these two properties. Let C_n be the number of regular expressions of length $2n$ and set $C_0 = 1$ for convenience. Show that

$$C_n = \sum_{k=1}^n C_{k-1} C_{n-k} = \sum_{j=0}^{n-1} C_j C_{n-j-1} \quad (n \geq 1).$$

These are the *Catalan numbers* first described by Eugène Catalan in 1838. The widespread appearance of these numbers owes to the prevalence of this recurrence in applications.

32. *Continuation.* Identify (with $+1$ and) with -1 . Argue that a necessary and sufficient condition for an expression of length $2n$ to be regular is that the corresponding walk returns to the origin in $2n$ steps through positive (i.e., ≥ 0) values only and hence show that $C_n = \frac{1}{n+1} \binom{2n}{n}$.

33. Let u_{2n} denote the probability that a random walk returns to the origin at step $2n$, and let f_{2n} denote the probability that it returns to the origin for the *first time* at step $2n$. Show that $u_{2n} = \sum_{j=1}^n f_{2j} u_{2n-2j}$ and hence verify that $u_{2n} = (-1)^n \binom{-1/2}{n}$ and $f_{2n} = (-1)^{n-1} \binom{1/2}{n}$.

34. *Continuation.* Hence show that $u_0 u_{2n} + u_2 u_{2n-2} + \dots + u_{2n} u_0 = 1$.

35. Using the method of images show that the probability that $S_{2n} = 0$ and the maximum of S_1, \dots, S_{2n-1} equals k is the same as $P\{S_{2n} = k\} - P\{S_{2n} = 2k + 2\}$.

36. For $0 \leq k \leq n$, the probability that the first visit to S_{2n} takes place at epoch $2k$ equals $P\{S_{2k} = 0\} P\{S_{2n-2k} = 0\}$.

37. *Maxima of random walks.* Let M_n be the maximum of the values S_0, S_1, \dots, S_n . Show that $P\{M_n = r\} = P\{S_n = r\} + P\{S_n = r+1\}$ for each $r \geq 0$.

38. *Continuation.* Show that for $i = 2k$ and $i = 2k + 1$, the probability that the walk reaches M_{2n} for the first time at epoch i equals $\frac{1}{2} P\{S_{2k} = 0\} P\{S_{2n-2k} = 0\}$.

IX

The Essence of Randomness

In casual language we equate randomness with uniform choice. Laplace would agree. In his treatise *Essai Philosophique sur les Probabilités* published in 1814 (as a populist version of his masterpiece *Théorie Analytique des Probabilités* which, published in 1812, enjoyed immediate critical success) he wrote:

The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favourable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favourable cases and whose denominator is the number of all the cases possible.

On philosophical grounds Laplace favoured placing a uniform prior on parameters of a chance experiment that are unknown. He goes on to write:

One regards two events as equally probable when one can see no reason that would make one more probable than the other, because, even though there is an unequal possibility between them, we know not which way, and this uncertainty makes us look on each as if it were as probable as the other.

We have seen the principle in action in Laplace's law of succession (Section II.8). In Bayesian terms, Laplace was inclined to place a uniform prior on indeterminate causes.

Random or uniform choice over a finite number of alternatives segues naturally into the uniform density over a continuum of choice in a bounded interval. But uniform choice fails as a model of randomness in unbounded settings as there is no reasonable mechanism for allowing equal chances over an unbounded range. The key here is the continuous analogue of the memoryless property that the reader has seen exhibited in discrete settings in the case of the geometric distribution. The exponential density, viewed as the limiting case of the geometric distribution, inherits the memoryless property in the continuum and this is precisely the feature that makes it so compelling as a model of randomness over the unbounded half-line.

C 1, 5, 8–10
A 2–4, 6, 7

The uniform and the exponential densities together inherit the mantle of Laplacian uncertainty and serve as natural models of randomness in the continuum. The reader will not in retrospect find it surprising that these fundamental densities are connected at a variety of levels and share a prominent and important rôle in the theory.

1 The uniform density, a convolution formula

As we saw in Example I.7.7 and again in Example VII.3.1, the uniform density arises naturally by consideration of an abstract game where each sample point corresponds to the *gedanken* experiment of an unending sequence of tosses of a coin. Formally, the *uniform density* on the unit interval is defined by $u(t) = 1$ if $0 < t < 1$ and $u(t) = 0$ otherwise. This density is also called *rectangular* on account of the shape of its graph. The distribution function $U(t)$ corresponding to this density is piecewise linear, with $U(t) = 0$ if $t \leq 0$, $U(t) = t$ if $0 < t < 1$, and $U(t) = 1$ if $t \geq 1$. Straightforward integration shows that the mean of the uniform density $u(t)$ is $1/2$ while its variance is $1/12$.

A random variable X taking values in the unit interval in accordance with the uniform density u is said to be *uniformly distributed in the unit interval*. The probability that X takes values in any subinterval (s, t) of the unit interval is given by $\int_s^t u(\alpha) d\alpha = t - s$. In other words, the probability of any subinterval of the unit interval is identified with the *length* (or *Lebesgue measure*) of the subinterval. This is the formal justification of the intuitive assignment of probabilities in Example I.7.7.

Extensions of the model to situations where the random variable X takes values in an interval (a, b) are straightforward. The probability that X takes values in a subinterval (s, t) is now proportional to the length of the subinterval, the proportionality constant so chosen that the interval (a, b) has probability one. It follows that $P\{s < X < t\} = (t - s)/(b - a)$ for every choice of $a \leq s < t \leq b$ and we say that X is *uniformly distributed in the interval* (a, b) . The corresponding density of X is given by $u_{a,b}(t) = (b - a)^{-1}$ for $a < t < b$. Of course, $u_{a,b}(t) = \frac{1}{b-a} u\left(\frac{t-a}{b-a}\right)$ so that all uniform densities belong to the same type. It follows quickly, *vide* our observations in Section VII.4, that $u_{a,b}(t)$ has mean $(a + b)/2$ and variance $(b - a)^2/12$. It is an instructive exercise to consider how the passage to the general uniform continuous model may be achieved through an appropriate sequence of discrete models.

EXAMPLES: 1) *Aloha: contention resolution and backoff.* The Aloha protocol was developed *circa* 1970 to provide radio communication between a central computer facility and various terminals scattered around the campuses of the University of Hawaii. In general, the protocol facilitates a broadcast communication capability where several users have access to a common broadcast radio channel such as a satellite up-link. In the simplest versions of the protocol,

each user attempts to broadcast information into the channel (for instance, to a satellite for subsequent relay) as soon as the information is available. Unfortunately, this greedy individual mode of transmission can result in collisions when two or more users wish to transmit at the same time. In order to resolve the collisions systematically and fairly the protocol requires colliding users to back off and retransmit after a random time interval. The retransmission times of n colliding users correspond to our problem with n independent, uniformly distributed random variables.¹

2) *When lightning strikes twice.* Our next example has a whimsical flavour. A rod of unit length is struck twice by lightning breaking it into three pieces. What is the probability that the three segments can be arranged so as to form a triangle? The reader may find it illuminating to make a guess before diving into the analysis.

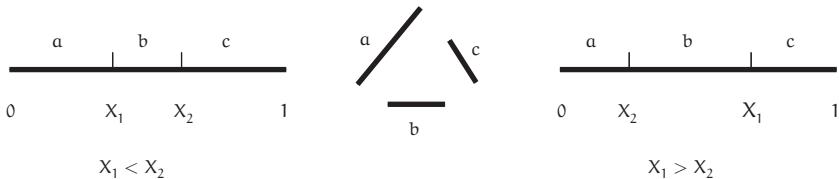


Figure 1: Possible breakage modes.

The key to the analysis is the following characteristic property of triangles: three line segments of lengths a , b , and c can be arranged to form a triangle if, and only if, the conditions $a + b > c$, $a + c > b$, and $b + c > a$ are satisfied simultaneously.

Suppose now that, as shown in Figure 1, the two lightning strikes occur at positions X_1 and X_2 on the rod and break the rod into three segments of lengths a , b , and c . Model X_1 and X_2 as independent random variables, each distributed uniformly in the unit interval. There are two symmetric cases possible, $X_1 < X_2$ and $X_1 > X_2$. If $X_1 < X_2$ then $a = X_1$, $b = X_2 - X_1$, and $c = 1 - X_2$ so that the necessary and sufficient conditions for the broken segments to form a triangle are $X_2 > 1/2$, $X_2 < X_1 + 1/2$, and $X_1 < 1/2$. If $X_1 > X_2$ then $a = X_2$, $b = X_1 - X_2$, and $c = 1 - X_1$ so that

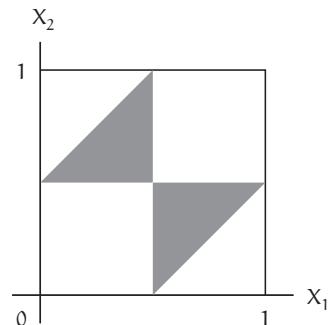


Figure 2: Sample points supporting triangular configurations.

¹The original version of the Aloha protocol used uniformly distributed backoff times. This has been superseded by geometric or exponential backoffs.

the necessary and sufficient conditions become $X_1 > 1/2$, $X_1 < X_2 + 1/2$, and $X_2 < 1/2$.

The region in the unit square satisfying these conditions corresponds to the bow-tie area shown shaded in Figure 2. The probability of the event Δ that the three segments can be arranged to form a triangle is hence equal to the probability that the random pair (X_1, X_2) takes values in the bow-tie region. It follows that

$$P(\Delta) = \iint_{\text{bowtie}} f(x_1, x_2) dx_2 dx_1$$

where $f(x_1, x_2)$ is the density of the pair (X_1, X_2) .

Now each of the X_i has a (marginal) distribution uniform in the unit interval with corresponding (marginal) density $f_i(x_i) = 1$ for $0 < x_i < 1$. As X_1 and X_2 are independent, it follows that the density $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ is identically equal to 1 in the unit square and is zero outside and, in consequence,

$$P(\Delta) = \iint_{\text{bowtie}} dx_2 dx_1 = 1/4$$

as the area of the bow-tie is exactly one-quarter of the area of the unit square. It follows that a double lightning strike will result in a potential triangle 25% of the time. ►

A little thought shows that a sum of uniformly distributed variables cannot also be uniform. Indeed, suppose X_1, X_2, \dots are independent and uniformly distributed in the unit interval. Begin with a consideration of the random variable $S_2 = X_1 + X_2$. If we write $u(x)$ for the uniform density in the unit interval, the density $u_2(x)$ of S_2 is given by

$$u_2(x) = \int_0^x u(t)u(x-t) dt = \begin{cases} x & \text{if } 0 < x \leq 1, \\ 2-x & \text{if } 1 < x < 2. \end{cases} \quad (1.1)$$

(Of course, $u_2(x)$ is identically zero elsewhere.) The graph of u_2 is a triangle and hence densities of this form are called *triangular*.

A similar exercise shows that the density $u_3(x)$ of the sum $S_3 = X_1 + X_2 + X_3$ is comprised of three quadratic segments in the interval $0 < x < 3$. And the explicit expressions for the densities get progressively more involved as the number of terms in the sum increases: the density $u_n(x)$ of $S_n = X_1 + \dots + X_n$ has support in the interval $0 < x < n$ and is comprised of n polynomial segments of degree $n - 1$ interspersed in this interval. Nonetheless, a systematic enumeration allows us to write down a general formula for the n -fold convolution of the uniform density.

For any real x it will be convenient to introduce the notation x_+ for the *positive part* of x defined by $x_+ = x$ if $x > 0$ and $x_+ = 0$ if $x \leq 0$. When n is even the expressions $(x_+)^n$ and $(x^n)_+ = x^n$ are not the same. To keep

from overburdening the notation, we adopt the slightly ambiguous short-hand notation x_+^n to mean $(x_+)^n$. Flying in the face of the usual convention regarding powers to zero, it will also prove useful to extend the definition to $n = 0$ by setting x_+^0 to be identically zero when $x \leq 0$ and equal to 1 when $x > 0$.

THEOREM Suppose X_1, \dots, X_n are independent random variables with common distribution uniform in the unit interval $(0, 1)$. Let $S_n = X_1 + \dots + X_n$ and write u_n for the density of S_n . Then

$$u_n(x) = \frac{1}{(n-1)!} \sum_{k=0}^n (-1)^k \binom{n}{k} (x-k)_+^{n-1}. \quad (1.2)$$

PROOF: For $n = 1$ we recover the uniform density. We now proceed by induction. As $S_{n+1} = S_n + X_{n+1}$ is a sum of independent variables, we have

$$u_{n+1}(x) = \int_0^\infty u_n(x-t)u(t) dt = \frac{1}{(n-1)!} \sum_k (-1)^k \binom{n}{k} \int_0^1 (x-t-k)_+^{n-1} dt, \quad (1.3)$$

by induction hypothesis. We may formally allow the sum to range over all integers k —our conventions for the binomial ensure that the only non-zero terms are for $0 \leq k \leq n$. The natural change of variable of integration $s \leftarrow x - t - k$ reduces considerations to an elementary integral and evaluating it we obtain

$$\begin{aligned} u_{n+1}(x) &= \frac{1}{n!} \sum_k (-1)^k \binom{n}{k} [(x-k)_+^n - (x-1-k)_+^n] \\ &= \frac{1}{n!} \sum_k (-1)^k \binom{n}{k} (x-k)_+^n - \frac{1}{n!} \sum_k (-1)^k \binom{n}{k} (x-1-k)_+^n. \end{aligned}$$

Replacing $k+1$ by k in the second term on the right and recombining terms via an application of Pascal's triangle results in

$$u_{n+1}(x) = \frac{1}{n!} \sum_k (-1)^k \left[\binom{n}{k} + \binom{n}{k-1} \right] (x-k)_+^n = \frac{1}{n!} \sum_k (-1)^k \binom{n+1}{k} (x-k)_+^n,$$

completing the induction. ▶

The expression (1.2) for u_n is only formally defined for all x ; as $0 < S_n < n$, it follows that $u_n(x)$ evaluates to identically zero for $x \leq 0$ or $x \geq n$. In the latter case we discover an unlooked for refinement of (IV.1.5) to non-integer $x \geq n$.

The result has some curiosity value in its own right and occasionally appears in applications, usually disguised. We shall see examples in the following sections. The explicit form of the density can also be turned to advantage in analysing the limiting behaviour of random sums. As in the case of the binomial, the limit theorems derived in the setting of the uniform distribution presage powerful general results of wide applicability.

2 Spacings, a covering problem

The uniform density serves as a model for *random points in an interval*. Suppose X_1, \dots, X_n are random variables generated by independent sampling from the uniform distribution in the unit interval. As the probability of coincidence of two random points is zero, with probability one the random points X_1, \dots, X_n partition the unit interval into $n + 1$ subintervals that we shall call the *spacings*. What can be said about the distribution of the lengths of these spacings? It may appear that the end spacings, anchored as they are by the fixed points 0 and 1, are in some way different from the interior spacings but, surprisingly, this is not so.

Let us begin with a single point (Figure 3). The point X_1 partitions the unit interval into two subintervals of lengths $L_1 = X_1$ and $L_2 = 1 - X_1$. It is clear that L_1 is uniformly distributed and, for reasons of symmetry, so is L_2 . Indeed, $P\{L_2 > t\} = P\{X_1 < 1 - t\} = 1 - t$ and, of course, $P\{L_1 > t\} = 1 - t$ as well. It follows that L_1 and L_2 have the same (uniform) distribution on the unit interval.

Two points X_1 and X_2 partition the unit interval into three subintervals of lengths, say, L_1 , L_2 , and L_3 . The situation with $X_1 < X_2$ is shown in Figure 4, though it is clear that an analogous situation prevails with $X_2 < X_1$ as well. Events in this probability space correspond to regions in the unit square defined by the coordinate variables X_1 and X_2 .

Thus, when $X_1 < X_2$ we have $L_1 = X_1$, $L_2 = X_2 - X_1$, and $L_3 = 1 - X_2$ and conditioned on the event $\{X_1 < X_2\}$ (geometrically the upper left triangle in the unit square) the event $\{L_1 > t\}$ corresponds to the set of sample points for which $\{X_1 > t\}$; likewise, under this conditioning, the events $\{L_2 > t\}$ and $\{L_3 > t\}$ may be identified with the set of points for which $\{X_2 - X_1 > t\}$ and $\{X_2 < 1 - t\}$, respectively. Of course, a completely analogous

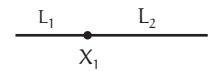


Figure 3: One random point.

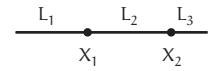


Figure 4: Two random points.

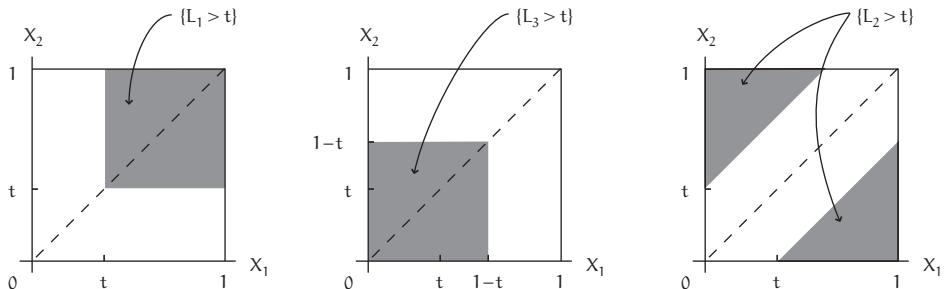


Figure 5: Distribution of spacings.

situation arises when $X_2 < X_1$, leading to a symmetric set of regions in the

lower right triangle in the unit square as seen in the shaded regions in Figure 5.

It is clear at once that the situation is completely symmetric and the three areas are all equal to the area of a square of side $1 - t$. It follows that the three spacings L_1 , L_2 , and L_3 all conform to the same distribution with $P\{L_j > t\} = (1 - t)^2$ for $j = 1, 2, 3$. The associated density has support in the unit interval and may be seen to be $2(1 - t)$ by differentiation.

The results for $n = 1$ and 2 suggest that we may, as a working hypothesis, adopt the proposition that, for any given n , all spacings have the same distribution. Events in this space correspond to regions inside the n -dimensional unit cube and the calculation of probabilities involves, at least formally, the computation of n -dimensional integrals over suitable regions. The tedium of the prospect does not appeal and it is worth casting about for an alternative attack that exploits the symmetry inherent in the problem.

Identify the points 0 and 1 and “roll up” the unit interval to form a circle of unit circumference which we denote \mathbb{T} . Picking a point at random in the unit interval is then equivalent, from a formal point of view, to picking a point uniformly at random on the circle \mathbb{T} . In this light, the n random points X_1, \dots, X_n are now points on the circumference of the circle. The special rôle of the point $0/1$ may be co-opted by picking another point X_0 independently of the others on the circle and considering the location of X_0 to be the origin. Cutting the circle at X_0 then reduces the problem to the original version of n random points in the unit interval with the arcs on either side of X_0 corresponding to the two end spacings and with the remaining $n - 1$ arcs forming the interior spacings (Figure 6).

What makes this viewpoint so effective is that the symmetry of the situation is palpable on the circle. The $n + 1$ points X_0, X_1, \dots, X_n partition \mathbb{T} into $n + 1$ arcs whose lengths, enumerating clockwise from X_0 are L_1, L_2, \dots, L_{n+1} ; in this view, the j th spacing L_j is the separation between the $(j - 1)$ th and j th points encountered starting at X_0 and proceeding clockwise around the circle. By symmetry it is now clear that *the arc lengths all have a common distribution*. It only remains to determine what this distribution is. It suffices to consider the arc of length L_1 adjacent clockwise to X_0 . The event $\{L_1 > x\}$ occurs if, and only if, the arc length in the clockwise direction from X_0 to each of X_1, \dots, X_n exceeds x . It follows that *all the spacings have the common marginal distribution* $P\{L_1 > x\} = (1 - x)^n$ and corresponding marginal density $n(1 - x)^{n-1}$. But much more can be said.

As the uniform distribution on the circle is invariant with respect to translations, by virtue of the independent selection of the random points it is clear that the joint distribution of (L_1, \dots, L_{n+1}) is invariant with respect to permutations; systems of variables with this property are said to be *exchangeable*.

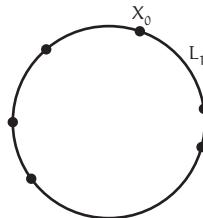


Figure 6: Spacings on the circle.

The symmetry inherent in the situation suggests that the joint distribution of the spacings may be determined exactly. And it can. We consider the problem in a slightly more general setting.

DE FINETTI'S THEOREM *Let $\tau > 0$ and suppose L_1, \dots, L_{n+1} are the successive spacings engendered by n random points in the interval $[0, \tau)$. Then*

$$P\{L_1 > x_1, \dots, L_{n+1} > x_{n+1}\} = \left(1 - \frac{x_1}{\tau} - \frac{x_2}{\tau} - \dots - \frac{x_{n+1}}{\tau}\right)_+^n$$

for every choice of $x_1 \geq 0, \dots, x_{n+1} \geq 0$.

PROOF: If $n = 1$ then L_1 is uniformly distributed in the interval $[0, \tau)$, $L_2 = \tau - L_1$, and so, if $x_1 \geq 0$ and $x_2 \geq 0$, we have

$$P\{L_1 > x_1, L_2 > x_2\} = P\{x_1 < L_1 < \tau - x_2\} = \frac{(\tau - x_1 - x_2)_+}{\tau} = \left(1 - \frac{x_1}{\tau} - \frac{x_2}{\tau}\right)_+.$$

This establishes the base of an induction.

Suppose now that $n > 1$. The event $L_1 > x$ occurs if, and only if, each of the random points X_1, \dots, X_n lies in the interval (x, τ) . Consequently, $P\{L_1 > x\} = (1 - \frac{x}{\tau})_+^n$ for $x \geq 0$. Differentiation shows that the marginal density of L_1 has support in $0 \leq x < \tau$ and is given by $\frac{n}{\tau}(1 - \frac{x}{\tau})^{n-1}$ for x in this interval.

Conditioned on $L_1 = x$, the spacings L_2, \dots, L_{n+1} are generated by $n-1$ random points in the interval $[0, \tau - x)$. And so, by induction hypothesis,

$$P\{L_2 > x_2, \dots, L_{n+1} > x_{n+1} \mid L_1 = x\} = \left(1 - \frac{x_2}{\tau - x} - \dots - \frac{x_{n+1}}{\tau - x}\right)_+^{n-1}$$

for any $x_2 \geq 0, \dots, x_{n+1} \geq 0$. Eliminate the conditioning by integrating out with respect to L_1 to obtain (total probability!)

$$\begin{aligned} P\{L_1 > x_1, L_2 > x_2, \dots, L_{n+1} > x_{n+1}\} \\ = \int_{x_1}^{\tau} P\{L_2 > x_2, \dots, L_{n+1} > x_{n+1} \mid L_1 > x\} \cdot \frac{n}{\tau} \left(1 - \frac{x}{\tau}\right)^{n-1} dx \\ = \frac{n}{\tau} \int_{x_1}^{\tau} \frac{(\tau - x - x_2 - \dots - x_{n+1})_+^{n-1}}{(\tau - x)^{n-1}} \cdot \frac{(\tau - x)^{n-1}}{\tau^{n-1}} dx. \end{aligned}$$

The change of variable $t = \tau - x - x_2 - \dots - x_{n+1}$ yields the elementary integral

$$P\{L_1 > x_1, L_2 > x_2, \dots, L_{n+1} > x_{n+1}\} = \frac{n}{\tau^n} \int_0^{(\tau - x_1 - x_2 - \dots - x_{n+1})_+} t^{n-1} dt$$

to complete the induction. ▶

The original elegant argument of de Finetti was essentially geometric and is not much longer.²

Specialising to the case of the unit interval, $\tau = 1$, de Finetti's theorem yields the joint spacing distribution,

$$\mathbf{P}\{L_1 > x_1, L_2 > x_2, \dots, L_{n+1} > x_{n+1}\} = (1 - x_1 - x_2 - \dots - x_{n+1})_+^n. \quad (2.1)$$

This distribution has support only in the n -dimensional simplex determined by the inequalities $x_1 \geq 0, \dots, x_{n+1} \geq 0$, and $x_1 + \dots + x_{n+1} \leq 1$. We can now determine the marginal distributions by setting one or more of the variables x_j to be zero. In particular, for any $x_j \geq 0$ and $x_k \geq 0$, we have

$$\mathbf{P}\{L_j > x_j\} = (1 - x_j)_+^n \quad (\text{for all } j), \quad (2.2)$$

$$\mathbf{P}\{L_j > x_j, L_k > x_k\} = (1 - x_j - x_k)_+^n \quad (\text{whenever } j \neq k), \quad (2.3)$$

and, in general, for any collection of k distinct indices j_1, \dots, j_k , we have

$$\mathbf{P}\{L_{j_1} > x_{j_1}, \dots, L_{j_k} > x_{j_k}\} = (1 - x_{j_1} - \dots - x_{j_k})_+^n \quad (2.4)$$

for all choices of positive x_{j_1}, \dots, x_{j_k} .

De Finetti's theorem is in repeated use in applications. Here is a *covering problem*. A collection of $n + 1$ arcs, each of length a , is thrown randomly on the circle \mathbb{T} of unit circumference. By random we mean that the centres of the arcs, say X_0, X_1, \dots, X_n are random points on the circle. What is the probability that the arcs cover the circle?

If the collection of $n + 1$ arcs does not cover the circle then there must exist a pair of neighbouring random points on the circle whose spacing exceeds $2 \cdot \frac{a}{2} = a$. Let A_j denote the event that the j th spacing L_j exceeds a . The event that the arcs do not cover the circle is hence given by the disjunction $A_1 \cup A_2 \cup \dots \cup A_{n+1}$.

De Finetti's theorem makes it simple to write down conjunction probabilities. For any collection of k distinct indices, j_1, \dots, j_k , by setting $x_{j_1} = \dots = x_{j_k} = a$ in (2.4), we obtain $\mathbf{P}(A_{j_1} \cap \dots \cap A_{j_k}) = (1 - ka)_+^n$. Summing over all choices of k indices, by exchangeability of the spacings, we have

$$S_k = \sum_{1 \leq j_1 < \dots < j_k \leq n+1} \mathbf{P}(A_{j_1} \cap \dots \cap A_{j_k}) = \binom{n+1}{k} (1 - ka)_+^n.$$

The stage is set for an application of the theorem of inclusion and exclusion [see (IV.1.1)]. Write $c_{n+1} = \mathbf{P}(A_1^c \cap \dots \cap A_{n+1}^c)$ for the probability that the arcs blanket the circle.

²B. de Finetti, *Giornale Istituto Italiano degli Attuari*, vol. 27, pp. 151–173, 1964.

THEOREM 2 *The probability that $n + 1$ random arcs, each of length a , cover the circle \mathbb{T} is given by*

$$c_{n+1} = \sum_{k=0}^{n+1} (-1)^k \binom{n+1}{k} (1 - ka)_+^n. \quad (2.5)$$

A comparison of (2.5) with (1.2) suggests that the two formulations are intimately related. Indeed, $c_{n+1} = a^n n! u_{n+1}(1/a)$. This cannot be accidental. And indeed it is not. Expand notation and write $c_{n+1}(\tau)$ for the probability that $n + 1$ random arcs cover a circle of circumference τ ; in this notation we identify $c_{n+1} = c_{n+1}(1)$ in the theorem. As in the proof of de Finetti's theorem, we condition on the first spacing. If $L_1 = x$, then the remaining spacings are generated by $n - 1$ points uniformly distributed in an interval of length $\tau - x$ and the probability that the corresponding arcs centred at these points cover the remaining real estate is, by definition, $c_n(\tau - x)$. As the first spacing must be covered by arcs surrounding X_0 and its first neighbour clockwise, we allow L_1 to vary between 0 and a . As we have seen, the density of L_1 is given by $\frac{n}{\tau} (1 - \frac{x}{\tau})^{n-1}$ and, by integrating out, we obtain the recurrence

$$c_{n+1}(\tau) = \int_0^\tau c_n(\tau - x) \cdot \frac{n}{\tau} \left(1 - \frac{x}{\tau}\right)^{n-1} dx.$$

For each n , define $\tilde{u}_n(t) = \frac{t^{n-1}}{(n-1)!} c_n(at)$. After some trivial algebraic consolidation, the recurrence then becomes

$$\begin{aligned} \tilde{u}_{n+1}(t) &= \frac{t^n}{n!} c_{n+1}(at) = \frac{t^n}{n!} \int_0^a c_n(a(t-x/a)) \cdot \frac{n}{at^n} \left(t - \frac{x}{a}\right)^{n-1} dx \\ &= \int_0^1 \frac{(t-y)^{n-1}}{(n-1)!} c_n(a(t-y)) dy = \int_0^1 \tilde{u}_n(t-y) dy. \end{aligned}$$

The convolutional recurrence (1.3) coyly peeks out of the fray and so the solution must be the same, $\tilde{u}_n = u_n$. The covering probability is hence related by a scale factor to the distribution of the sum of independent uniform variates.

An explicit formula is always worth having but it must be admitted that the formulation (2.5) is more opaque than most. We shall return to the problem from a fresh vantage point in Section XVIII.9 and discover a more succinct approximate characterisation.

3 Lord Rayleigh's random flights

In three dimensions a random point on the unit sphere connotes a random element drawn from the uniform distribution on the surface \mathbb{S}^2 of the ball of unit radius. This is the natural analogue in three dimensions of selecting a random

point on the unit circle. In this context, a happy dimensional accident makes calculations in three dimensions simple.

LEMMA 1 *The surface area of the portion of the sphere \mathbb{S}^2 that is contained within two parallel planes intersecting it depends only on the separation of the planes.*

PROOF: A point on the sphere is specified by its latitude ϑ and its longitude φ . For definiteness, we measure latitude with respect to the angle with the ray passing through the North Pole, $0 \leq \vartheta < \pi$, and longitude counter-clockwise along the equator with respect to any convenient longitude as reference, say the International Date Line, $0 \leq \varphi < 2\pi$. Suppose two parallel planes separated by a distance h intersect the sphere. We may suppose, without loss of generality, that the planes are perpendicular to the North-South axis, the intersection of these planes with the sphere at latitudes ϑ_1 and ϑ_2 , for definiteness, say, $\vartheta_1 < \vartheta_2$. Then $h = \cos \vartheta_1 - \cos \vartheta_2$. Now an infinitesimal area element on the sphere at coordinate location (ϑ, φ) has area $\sin \vartheta d\vartheta d\varphi$ and so the area of the sphere trapped between the planes is given by

$$\int_0^{2\pi} \int_{\vartheta_1}^{\vartheta_2} \sin \vartheta d\vartheta d\varphi = 2\pi(\cos \vartheta_1 - \cos \vartheta_2) = 2\pi h,$$

the result depending only on the separation h and not the particular latitudes of intersection ϑ_1 and ϑ_2 . ▶

LEMMA 2 (PROJECTION LEMMA) *Suppose \mathbf{X} is a random point on the sphere \mathbb{S}^2 . Then its projection, say, \mathbf{X}^0 , in any given direction along a ray passing through the origin, say, the North-South axis, is uniformly distributed in the interval $[-1, 1]$.*

This follows directly from the previous lemma. While it is tempting to imagine that the result codifies an intrinsic, dimension-independent property of projections of random points on a sphere, the result is false in two dimensions as the reader should be able to quickly verify. Our projection lemma is a dimensional artefact peculiar to three dimensions.

Now suppose $\mathbf{X}_1, \dots, \mathbf{X}_n$ are random points on the unit sphere \mathbb{S}^2 —the reader should think of these as unit vectors with random orientations representing movements, velocities, molecular links, and such like. Let $\mathbf{S}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$ be their vector sum. In applications \mathbf{S}_n can represent a long molecular chain, a random walk in three dimensions (or *random flight*), the vector accumulation of random perturbant velocities, or, in the original setting which led Karl Pearson to formulate this problem in July 1905, the spatio-temporal evolution of mosquito populations. We are indebted to Pearson for coining the evocative phrase “random walk”. In his article, Pearson wrote:³

³The quote is reproduced by permission from Macmillan Publishers Ltd: K. Pearson, “The problem of the random walk”, *Nature*, vol. 72, p. 294, 1905. Copyright © Nature Publishing Group, 1905.

A man starts from a point 0 and walks 1 yards in a straight line; he then turns through any angle whatever and walks another 1 yards in a second straight line. He repeats this process n times. I require the probability that after n of these stretches he is at a distance between r and $r + \delta r$ from his starting point.

Lord Rayleigh provided an answer in the following week by pointing out that Pearson's random walk question was subordinated in an earlier analysis of Rayleigh's on sound vibrations. I cannot resist the urge to quote once more from Pearson who wrote in reference to Rayleigh's letter in the August 1905 issue of *Nature*, "I ought to have known it ... [but] one does not expect to find the first stage of a biometric problem provided in a memoir on sound." Policy makers may look to this anecdote for support for a policy of broad general education on the principle that any two spheres of knowledge may be connected.

This is the setting. We are primarily interested in the Euclidean length $\|\mathbf{S}_n\|$ of the vector sum \mathbf{S}_n . Let v_n and V_n , respectively, represent the density and d.f. of $\|\mathbf{S}_n\|$. The length provides information about the degree of compaction or diffusion of the sum. If $\|\mathbf{S}_n\|$ is small then, in a biomolecular example, it may suggest a high degree of spatial compaction of a long molecule; if the sum is viewed as a three-dimensional random walk (or flight) then a small value for $\|\mathbf{S}_n\|$ suggests that the walk has remained in a proximity of the origin.

The ray from the origin passing through the point \mathbf{S}_n intersects the unit sphere \mathbb{S}^2 at the point $\Omega_n = \mathbf{S}_n / \|\mathbf{S}_n\|$. This is the unit vector in the direction of \mathbf{S}_n and hence represents the *orientation* of \mathbf{S}_n in space. By the symmetry of the situation, Ω_n is a random point in \mathbb{S}^2 and is independent of the length $\|\mathbf{S}_n\|$.

The analysis of the distribution of the length of the vector sum is surprisingly facilitated by considering its projection in any given direction, say, along the North-South axis for definiteness. Accordingly, let T_n denote the projection of \mathbf{S}_n along the North-South axis; likewise, let Z_n denote the projection of Ω_n , also in this direction. Then $T_n = Z_n \|\mathbf{S}_n\|$. Let w_n and W_n denote the density and d.f., respectively, of T_n .

By virtue of Lemma 2, as Ω_n is a random point on the sphere \mathbb{S}^2 independent of $\|\mathbf{S}_n\|$, its projection Z_n is uniformly distributed in the interval $[-1, 1]$ and is also independent of $\|\mathbf{S}_n\|$. By conditioning on the value of Z_n , we then have

$$W_n(t) = \int_{-1}^1 \mathbf{P}\{x \|\mathbf{S}_n\| \leq t\} \frac{dx}{2} = \frac{1}{2} \int_{-1}^0 \left[1 - V_n\left(\frac{t}{x}\right) \right] dx + \frac{1}{2} \int_0^1 V_n\left(\frac{t}{x}\right) dx.$$

And so, by differentiating under the integral sign, we obtain

$$w_n(t) = -\frac{1}{2} \int_{-1}^0 v_n\left(\frac{t}{x}\right) \frac{dx}{x} + \frac{1}{2} \int_0^1 v_n\left(\frac{t}{x}\right) \frac{dx}{x}.$$

As v_n has support only in the positive half-line, if $t > 0$ only the second term on the right survives, while if $t < 0$ only the first term is non-zero. By symmetry it

suffices to consider only one of the cases. Suppose accordingly that $t > 0$. Then

$$w_n(t) = \frac{1}{2} \int_0^1 v_n\left(\frac{t}{x}\right) \frac{dx}{x} = \frac{1}{2} \int_t^\infty v_n(y) \frac{dy}{y} \quad (t > 0)$$

by the indicated change of variable $y = t/x$ with differential element $-dy/y = dx/x$. By differentiating both sides, an appeal to the fundamental theorem of calculus shows that $w'_n(t) = -v_n(t)/(2t)$ or, equivalently,

$$v_n(t) = -2tw'_n(t) \quad (t > 0).$$

Here is a beautiful and unexpected relation connecting the length of a random vector sum of unit vectors to its projection. If we can somehow determine the distribution of the projection along any given axis then we have a pathway to determining the distribution of the length $\|S_n\|$. And in view of (1.2) a formula is at hand.

Write X_j^0 for the projection along the chosen direction of the point X_j on the sphere. We may then identify $T_n = X_1^0 + \dots + X_n^0$. But X_1^0, \dots, X_n^0 are independent and uniformly distributed in $[-1, 1]$ by Lemma 2 and their sum differs from that of a sum of random points in the unit interval by only a trivial shift of origin and a scale: explicitly, $S_n^0 = (T_n + n)/2$ connotes a sum of random points in the unit interval. As S_n^0 and T_n have densities of the same type, we may hence immediately write down the density of T_n as given by $w_n(t) = \frac{1}{2}u_n\left(\frac{t+n}{2}\right)$. Differentiating (1.2) termwise now yields v_n .

THEOREM *Suppose $S_n = X_1 + \dots + X_n$ is the vector sum of n random points on the sphere \mathbb{S}^2 . Then its length $\|S_n\|$ has density*

$$v_n(t) = \frac{-t}{2^{n-1}(n-2)!} \sum_{k=0}^n (-1)^k \binom{n}{k} (t+n-2k)_+^{n-2} \quad (t > 0).$$

These densities appear in a famous paper of the physicist S. Chandrasekhar.⁴ Using Markov's method (see Section VI.4), Chandrasekhar provided a general solution for the density $v_n(t)$ in the form of a Fourier integral and evaluated it explicitly (following Rayleigh) for $n = 3, 4$, and 6 ; the graphs of these densities appear in Figure 7. The reader who peruses Chandrasekhar's paper will be struck by how the use of the elementary projection lemma finesse the need for much difficult calculation—and provides an explicit general result to boot. As is frequently the case, the adoption of a proper perspective brings a magical clarity to a murky business.

⁴S. Chandrasekhar, "Stochastic patterns in physics and astronomy", in *Selected Papers on Noise and Stochastic Processes* (ed. N. Wax). New York: Dover, 1954. As Chandrasekhar worked in spherical coordinates his formulae need to be multiplied by $4\pi t^2$ to recover our expressions.

The projection lemma has utility in other settings. Audio waveforms arriving at a listener at a concert may be modelled as random unit vectors, the perceived intensity varying proportional to the *square* of the amplitude of the superposition of these waveforms. If two sources are separated by a random angle Θ the square of the length of the received vector is $2+2 \cos \Theta$. But the projection lemma tells us that $\cos \Theta$ is uniformly distributed in the interval $[-1, 1]$, hence has zero expectation. And so the expected value of the recorded intensity is indeed 2 even though the hearing mechanism involves a square. As W. Feller has pointed out, this explains why two violins are twice as loud as one.

As a postscript, beauty is indeed in the eye of the beholder. The recording of light waves, as for audio waves, involves the intensity or the square of the amplitude of a superposition of wavefronts. But depth perception is squirrelled away in the phase or angle. Thus, images obtained in ordinary photography are two-dimensional projections bereft of depth data. What is good for the audiophile is not so good for precise imaging. D. Gabor's Nobel Prize-winning work for the development of the principle of optical holography in 1947 showed how the phase information critical to depth perception could be stored and retrieved in photographic film even though the storage mechanism involves intensities and not amplitudes.

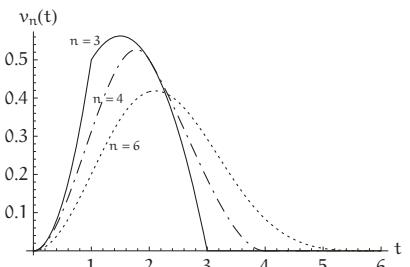


Figure 7: Rayleigh–Chandrasekhar densities.

4 M. Poincaré joue à la roulette

As we saw in the previous examples, it is sometimes fruitful to consider the sample space of the unit interval to be the circumference of a circle. In other applications, while the coordinate random variable X takes values on the real line the attribute of X of interest is just its fractional part. In such settings it is natural to consider the *reduced random variable* ${}^0X = X \bmod 1$ which is just the value of X reduced modulo 1 to its fractional part. The reduced variable 0X may be considered to be a random point on the circumference of a circle (of radius $1/2\pi$). The game of chance roulette provides a simple case in point.

EXAMPLE 1) Poincaré's roulette problem. Ian Fleming's intrepid protagonist James Bond made a habit of winning at roulette. But, in the absence of malign control by the gambling house, naïve intuition suggests that the number of gyrations of a roulette wheel results in a position uniformly distributed on the unit circle (which leaves a suspicious cast on Bond's unerring eye for a win). H. Poincaré put this vague intuition on a formal footing of a limit theorem in the year 1912.

We shall be satisfied here with a broad-brush analysis.

Following W. Feller, let X denote the distance covered in a single spin of the roulette wheel, $F(x)$ and $f(x)$ its d.f. and density, respectively; analogously, let ${}^0F(x)$ and ${}^0f(x)$ be the d.f. and density, respectively, of the reduced variable 0X . Then, for $0 \leq x < 1$,

$${}^0F(x) = \mathbf{P}\{{}^0X \leq x\} = \mathbf{P}\left(\bigcup_n \{n < X \leq n + x\}\right)$$

and countable additivity of probability measure allows us to conclude that

$${}^0F(x) = \sum_n \mathbf{P}\{n < X \leq n + x\} = \sum_n [F(n + x) - F(n)].$$

By a formal differentiation under the summation sign, we obtain ${}^0f(x) = \sum_n f(n + x)$.⁵

If 0X is to be approximately uniformly distributed then ${}^0f(x) \approx 1$ for all $0 < x < 1$. Under what conditions on f would this be true? Consider for instance a unimodal density $f(t)$ with a maximum value m achieved at $t = \xi$ (Figure 8).

The intervals $[k + \xi, k + 1 + \xi]$ as k ranges over the integers partition the real line and, for every fixed x , each such interval $k + \xi \leq t < k + 1 + \xi$ contains precisely one point of the form $n + x$ for integer n . Write $t_k = n + x$ to make explicit the containing interval. Then $f(n + x) = f(t_k) = \int_{k+\xi}^{k+1+\xi} f(t) dt$. It follows that

$$\begin{aligned} |{}^0f(x) - 1| &= \left| \sum_n f(n + x) - \int f(t) dt \right| = \left| \sum_k \int_{k+\xi}^{k+1+\xi} f(t_k) dt - \sum_k \int_{k+\xi}^{k+1+\xi} f(t) dt \right| \\ &= \left| \sum_k \int_{k+\xi}^{k+1+\xi} (f(t_k) - f(t)) dt \right| \leq \left| \sum' \right| + \left| \sum'' \right| \end{aligned}$$

where \sum' sums over $k < 0$ and \sum'' sums over $k \geq 0$. As $f(t)$ decreases monotonically with t for $t \geq \xi$, we have

$$\left| \sum'' \right| \leq \sum_{k \geq 0} \int_{k+\xi}^{k+1+\xi} |f(t_k) - f(t)| dt \leq \sum_{k \geq 0} [f(k + \xi) - f(k + 1 + \xi)] = f(\xi) = m$$

as the sum on the right telescopes. Similarly, $f(t)$ increases monotonically with t for $t < \xi$ and a similar argument shows that $|\sum'| \leq m$. It follows that $|{}^0f(x) - 1| \leq 2m$.

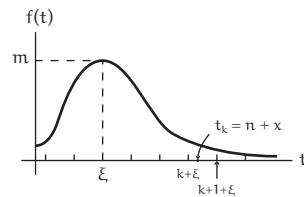


Figure 8: Unimodal density for the distance covered in one spin of the roulette wheel.

⁵If the reader is worried about convergence of the formal series for 0f , it will suffice for her to assume that f decreases monotonically in the tails as the convergence of the integral $\int f(t) dt$ will then imply the convergence of the series $\sum_n f(n + x)$.

$|f'(x)| \leq 2m$ for all $0 < x < 1$ and, in consequence, ${}^0f(x)$ is approximated by the uniform density if m is small. The reader will have realised that the unimodality of f is not critical here. *The reduced variable 0X will be approximately uniformly distributed in the unit interval if X has a density that is sufficiently spread out.* ►

The reduced random variable 0X hence shows a tendency to the uniform (though we shall not try to make this more precise at this juncture). This kind of phenomenon also makes its appearance in other settings where it is rather less expected.

EXAMPLE 2) *Benford's law of anomalous numbers.* A random page is selected from a voluminous compendium such as the *Farmer's Almanac* or, given the volubility of politicians, perhaps a *Congressional Record*, by a neutral party. Your favourite card sharp offers you 1-for-1 odds that the most significant digit of the page number unearthed is less than 5. (In other words, you receive \$1 from the card sharp if the most significant digit is 5 or larger; you pay up \$1 if the most significant digit is 4 or less.) Are these good odds?

On the face of it, it appears that the card sharp has made a mistake. One may naturally expect that each of the digits from 1 through 9 has an equal chance of being the most significant digit. This would imply that the probability that the most significant digit is 1, 2, 3, or 4 is $4/9$ which is less than $1/2$. If the reader plays the game at these odds n times her mean winnings should then be $n/9$ and the card sharp should rapidly go out of business.

In 1938 F. Benford provided convincing empirical evidence, however, that this is not the case: the odds of the most significant digit being less than 5 are approximately 70%, not 44%.⁶ And the card sharp laughs all the way to the bank. How to read this riddle? If, for instance, the compendium contains 9999 pages, then each of the digits from 1 through 9 appears as the most significant digit in exactly 1111 pages so that each digit is the most significant precisely one-ninth of the time. The fallacy in our thinking occurs at an unexpected place and we now proceed to explain it following a line of heuristic reasoning credited to R. S. Pinkham by W. Feller.

If the compendium is large, sampling will rarely approximate a uniform distribution. Suppose the page number Z that is drawn conforms to some unspecified distribution. To obviate trivialities we suppose that the distribution of Z is spread out over a large range, hence the requirement of a *voluminous* compendium. It is useful to think of the distribution of Z as an approximation of a distribution with support on the positive integers. In particular, this rules out the uniform distribution. The most significant digit of Z is k if, and only if, $10^n k \leq Z < 10^n(k+1)$ for some positive integer n . If we write $X = \log_{10} Z$, this is equivalent to saying that $n + \log_{10} k \leq X < n + \log_{10}(k+1)$. The integer

⁶F. Benford, "The law of anomalous numbers", *Proceedings of the American Philosophical Society*, vol. 78, pp. 551–572, 1938.

offset n is not relevant here and what matters is the fractional part of X , that is to say, the reduced random variable ${}^0X = X \bmod 1$. It follows that the most significant digit is k if, and only if, $\log_{10} k \leq {}^0X < \log_{10}(k+1)$. But, following Poincaré, the distribution of the reduced variable 0X is approximately uniform if the distribution of Z is sufficiently spread out. It follows that the probability that the most significant digit is k is approximately $p_k = \log_{10}(k+1) - \log_{10} k$. These numbers (truncated to three decimal places) are listed in Table 1.

p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Table 1: The distribution of significant digits.

The probability that the most significant digit is less than 5 is then approximately $p_1 + p_2 + p_3 + p_4 = \log_{10} 5$, which is about 0.7, and the approximation will be good if Z is spread out over a large range. The fault in our original intuition is in the naïve expectation that the obvious uniformity of the most significant digit when the sampling distribution is uniform will carry over when the sampling distribution deviates from the uniform as it is very likely to do for a large compendium. In this case, the uniform distribution does arise again, but it is the reduced log variable that is approximately uniform. ▶

A classical application where the reduced variable 0X makes a natural appearance provided an early empirical estimate of π .

EXAMPLE 3) *Buffon's needle problem.* A unit-length needle is tossed and lands at a random position on the plane on which is scored an (infinite) family of parallel lines, one unit apart, and extending to infinity in both directions as illustrated in Figure 9. What is the probability that the needle intersects any of the parallel lines?

We may model the situation as corresponding to the specification of two random variables, the horizontal position X of the centre of the needle and the angle $\Theta \in (-\pi/2, \pi/2)$ that it makes with respect to the positive x -axis. More precisely, the positional random variable of interest is in actuality the reduced variable ${}^0X = X \bmod 1$ which specifies the distance of the centre of the needle from the nearest grid line to the left. While the distribution of the pair $({}^0X, \Theta)$ is not *a priori* specified by the experiment, on intuitive grounds, we may suppose that 0X and Θ are independent and uniformly distributed in the intervals $(0, 1)$ and $(-\pi/2, \pi/2)$, respectively.

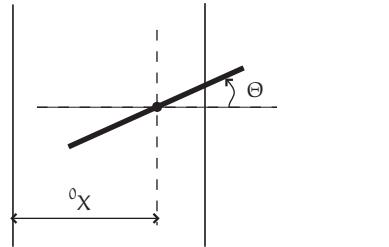


Figure 9: Buffon's needle in the plane.

Let \mathbb{A} be the event that the needle crosses a line. If \mathbb{A} occurs then precisely one of the following two events has to have occurred: either the left half of the needle crosses a vertical line or the right half of the needle does. It follows that \mathbb{A} is comprised of the set of sample points $(^0x, \theta)$ for which $0 \leq {}^0x < \cos(\theta)/2$ or $1 - \cos(\theta)/2 \leq {}^0x < 1$, shown shaded in Figure 10. These events are mutually exclusive and so, integrating out over the region \mathbb{A} , we obtain

$$\begin{aligned} P(\mathbb{A}) &= \iint_{\mathbb{A}} f({}^0x, \theta) d{}^0x d\theta \\ &= \int_{-\pi/2}^{\pi/2} \int_0^{\frac{1}{2} \cos \theta} \frac{dxd\theta}{\pi} + \int_{-\pi/2}^{\pi/2} \int_{1-\frac{1}{2} \cos \theta}^1 \frac{dxd\theta}{\pi} = \frac{1}{\pi} \int_{-\pi/2}^{\pi/2} \cos \theta d\theta = \frac{2}{\pi}. \end{aligned}$$

This calculation was originally performed by the Comte de Buffon in 1777 and, as he observed, this provides an empirical method to determine π to any required accuracy. Let S_n denote the number of line crossings in n independent tosses of a needle. For large enough values of n the weak law of large numbers [see Section V.6 and Problem V.17; the general formulation is provided in Chapter XVI] tells us that the relative frequency of line crossings, S_n/n , approximates $P(\mathbb{A})$ with high confidence. The ratio $2n/S_n$ hence provides us a principled, probabilistic, approximation for π . In 1850, the astronomer R. Wolf tossed a needle 5000 times in Zurich and obtained the estimate 3.1596 for π (instead of $3.14159\dots$).

How many tosses do we need to guarantee the first two digits in the decimal expansion of π with a confidence of 99%? What if we want the first five digits to the same level of confidence? ▶

One may query whether the accuracy in the estimate for π can be improved by altering the nature of the experiment. Exercises in this vein are sketched in the *Problems* at the end of this chapter.

The examples above vividly illustrate the characteristic property at the heart of the uniform distribution: *probabilities of regions depend only on their length (area, volume) and not on their location or shape*. In particular, this implies a translation invariance. Such symmetries lurk at the heart of problems derived out of the uniform and are key to their efficient resolution.

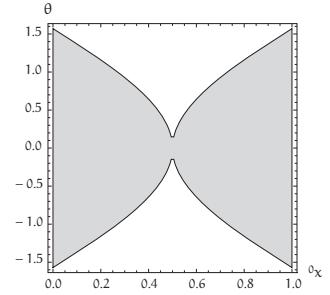


Figure 10: Buffon crossings.

5 Memoryless variables, the exponential density

As we saw in Section VIII.7, the geometric distribution is the unique arithmetic distribution featuring the discrete memoryless property. As the exponential

density of Example VII.3.2 is the continuous limit of a sequence of geometric distributions, it is worthy of investigation to see whether the passage from a discrete to a continuous model preserves this property.

For given $\alpha > 0$, the *exponential density with parameter α* is given by $f(t) = \alpha e^{-\alpha t}$ for $t > 0$. Corresponding to it is the continuous distribution function $F(t) = 1 - e^{-\alpha t}$ concentrated on the positive half-axis $t > 0$. A continuous-valued random variable X with a density of this form is said to be *exponentially distributed with parameter α* . As is easy to see, $\int_{-\infty}^{\infty} tf(t) dt = \int_0^{\infty} \alpha te^{-\alpha t} dt = \frac{1}{\alpha}$ so that $1/\alpha$ may be identified with the *mean*. It is hence also permissible and indeed customary to say that X is *exponentially distributed with mean $1/\alpha$* . Two integrations by parts show likewise that the variance of X is $1/\alpha^2$.

It is clear that all exponential distributions are of the same type. Indeed, by comparison with the unit mean exponential density $g(t) = e^{-t}$ it is clear that any other exponential density $f(t) = \alpha e^{-\alpha t}$ is of the form $\alpha g(\alpha t)$.

It is now easy to verify that the discrete memoryless property of the geometric distribution has a natural counterpart in the continuum. Let X be a continuous-valued random variable whose distribution has support in the positive half-axis. We say that X exhibits the *memoryless property* (or more succinctly that X , or its distribution, is *memoryless*) if $P\{X > s + t | X > s\} = P\{X > t\}$ for any choices of $s, t \geq 0$. If X denotes the waiting time before the occurrence of some event (the time between customer arrivals in a queue or between photon emissions in a low light setting to take two examples at random) then the memoryless property says that the elapsed time does not influence the future waiting time; in other words, the system does not retain memory of its past. It is now easy to see that the exponential distribution inherits its characteristic memoryless property from the geometric distribution.

The occurrence of the event $\{X > s + t\}$ clearly implies the occurrence of the event $\{X > s\}$ and so $P\{X > s + t | X > s\} = P\{X > s + t\}/P\{X > s\}$. We may hence recast the memoryless property in the simple form

$$P\{X > s + t | X > s\} = P\{X > t\} \text{ if, and only if, } P\{X > s + t\} = P\{X > s\} P\{X > t\} \quad (5.1)$$

for every $s, t \geq 0$. If X is exponentially distributed with parameter α then $P\{X > x\} = e^{-\alpha x}$ for all positive x and the condition (5.1) is easily seen to hold. It follows that the exponential distribution is indeed memoryless. We can go a little further: *a continuous distribution with support in the positive half-axis is memoryless if, and only if, it is exponential*.

The proof that the exponential density is the unique density possessed of the memoryless property is elementary and a consequence of the characteristic product property of the exponential function. Suppose X is any positive random variable possessed of a continuous distribution exhibiting the memoryless property. Write $G(t) = P\{X > t\}$. Some immediate facts about G may be deduced. As X is positive, $G(t) = 1$ for $t < 0$ and as the distribution is assumed continuous it follows that $G(0) = 1$ as well. Continuity of probability measure also ensures that $G(t)$ is monotone with

$G(t) \rightarrow 0$ as $t \rightarrow \infty$. Now for the particular requirements on G specific to the problem: the memoryless condition (5.1) implies that G must satisfy the functional equation $G(s+t) = G(s)G(t)$ for all positive s and t .

Write $\gamma(t) = \log G(t)$. Then γ inherits continuity and monotonicity from G , $\gamma(t) = 0$ for $t \leq 0$, and $\gamma(t) < 0$ for $t > 0$ with $\gamma'(t)$. The memoryless property is now equivalent to $\gamma(s+t) = \gamma(s) + \gamma(t)$ for all $s, t \geq 0$.

Suppose $\gamma(1) = -\alpha$ for some positive α . Then $-\alpha = \gamma(\frac{1}{2} + \frac{1}{2}) = 2\gamma(\frac{1}{2})$ and by induction, $-\alpha = \gamma(\frac{1}{n} + \dots + \frac{1}{n}) = n\gamma(\frac{1}{n})$ for every integer $n \geq 1$. It follows that $\gamma(n^{-1}) = -\alpha n^{-1}$. Observe that continuity at the origin is preserved as $\gamma(n^{-1}) \rightarrow 0$ as $n \rightarrow \infty$. Now $[\gamma(\frac{1}{n}) - \gamma(0)]/\frac{1}{n} = -\alpha$ and taking the limit as $n \rightarrow \infty$ of both sides we observe that γ has a derivative from the right at the origin given by $\gamma'(0+) = -\alpha$. Now, in the neighbourhood of any $t > 0$, for every $0 < \epsilon \leq t$, we have $\gamma(t+\epsilon) - \gamma(t) = \gamma(\epsilon)$ and, likewise, $\gamma(t-\epsilon) - \gamma(t) = -\gamma(\epsilon)$. It follows that $\epsilon^{-1}(\gamma(t+\epsilon) - \gamma(t)) = \epsilon^{-1}\gamma(|\epsilon|)$ for all sufficiently small ϵ . Letting $\epsilon \rightarrow 0$ on both sides we discover that γ is differentiable at t with $\gamma'(t) = \gamma'(0+) = -\alpha$. By continuity, $\gamma(0+) = \gamma(0) = 0$, and it follows that $\gamma(t) = -\alpha t$ for all $t > 0$ or, what is the same thing, $G(t) = e^{-\alpha t}$ for $t > 0$. It follows that if a continuous distribution with support in the positive half-axis exhibits the memoryless property then it must be exponential.

EXAMPLE: *The weakest link.* In a continuous analogue of the weakest link in a chain we are interested in the point on a thread or string that is most susceptible to breakage. Applications include long-chain polymers, yarn, balls of twine, and threads of fabric. A point of breakage will not lie athwart of two adjacent strands of lengths s and t , respectively, if, and only if, both strands individually bear up to the load. If X denotes the point of first breakage this says that $P\{X > s+t\} = P\{X > s\}P\{X > t\}$ if we assume no interaction between the strands. It follows that X has an exponential distribution. ▶

The memoryless property is a characteristic and uniquely defining property of the exponential distribution. This is the reason why the exponential distribution is so prominent in the theory of Markov processes (in which context the memoryless property is referred to as the *Markov property*) and in queuing theory. *Just as the translation invariance property is key to the uniform density, the memoryless property is central to the exponential density and may be found lurking at the heart of problems involving it.*

6 Poisson ensembles

The key point behind the memoryless property of the exponential distribution is that non-overlapping intervals do not interact. This basic feature can make its appearance in spaces that have more than one dimension.

EXAMPLE 1) *Ensembles of stars.* Suppose stars are distributed at random in an infinite space. On the scale of interstellar distances in units of, say, light years,

the stars are to be thought of as point masses. With this proviso it is natural to consider that the ensemble of stars has the following two properties: (i) the number of stars in any given region depends only on the volume of the region (say in parsecs, the volume of a cube one light year a side); and (ii) the number of stars in a region is independent of the number in any non-overlapping region.

Let \mathbb{A} and \mathbb{B} be two non-overlapping regions of volumes v and w , respectively. Write A and B for the events that the regions \mathbb{A} and \mathbb{B} are empty, respectively. Then, by our independence hypothesis, $P(A \cap B) = P(A)P(B)$. Write $G(v)$ for the probability that a region of volume v is empty. Then $A \cap B$ is the event that a region of total volume $v + w$ is empty, hence has probability $G(v + w)$. It follows that $G(v + w) = G(v)G(w)$ for all $v, w > 0$ and in consequence there is a positive α for which $G(v) = e^{-\alpha v}$ for $v > 0$.

Thus, the probability that a ball of radius r in space is devoid of stars is $e^{-4\alpha\pi r^3/3}$. It should be clear that the geometry of the underlying space plays no essential role here. Thus, if we are dealing with a two-dimensional space then area replaces volume and the probability that there is no “star” within a circle of radius r is $e^{-\alpha\pi r^2}$. Likewise, on the real line, length replaces volume and the probability that an interval of length $2r$ is empty is $e^{-2\alpha r}$. More generally, the Euclidean ball $\mathbb{B}^n(r)$ of radius r and centred at the origin in n dimensions is defined to be the collection of points (x_1, \dots, x_n) for which $x_1^2 + \dots + x_n^2 \leq r^2$. The volume of such a ball is given by the n -dimensional integral $\text{Vol } \mathbb{B}^n(r) = \int \dots \int_{\mathbb{B}^n(r)} dx_n \dots dx_1$ which, by the change of variable $x_j \leftarrow x_j/r$, we may express as $\text{Vol } \mathbb{B}^n(r) = r^n \text{Vol } \mathbb{B}^n(1) = r^n V_n$ where V_n is the volume of the unit ball in n dimensions. Thus, if we are dealing with a random ensemble of points in Euclidean n -space the probability that a ball of radius r is empty is given by $e^{-\alpha r^n V_n}$. The nature of the dependence of V_n on n is not important for our purposes here but the reader who has a spare afternoon can determine an explicit expression for V_n by a transformation to spherical coordinates—or, if she is short of time and patience, she can simply refer to Section XIV.6. ▶

A high-dimensional joke: string theory in physics would have us inhabiting a high-dimensional universe with many tiny (coiled!) dimensions adding pizzazz to the visible structure of three-dimensional space. Balls in such spaces are presumably not Euclidean but, with a suitable definition of volume, our theory will work unabated for random ensembles of “stars” in such spaces.

EXAMPLE 2) *The distribution of stars.* What can be said about the distribution of the *number* of stars? Write $N(v)$ for the number of stars in any given region of volume v and, for each integer $n \geq 0$, write $p_n(v) = P\{N(v) = n\}$. In this notation, we’ve just shown that $p_0(v) = G(v) = e^{-\alpha v}$. The general form of the probabilities $p_n(v)$ may now be determined via a perturbation argument introduced by M. W. Crofton in the late nineteenth century.

The basic idea behind *Crofton's method* is to consider the situation when we perturb a given volume by an infinitesimal amount. With high probability, the infinitesimal added volume will contain at most one point and any such point is effectively “anchored” in a tiny space allowing a recurrence to be set up for the remaining points.

To formalise, the probability that an infinitesimal region of volume ϵ is empty is given by $p_0(\epsilon) = e^{-\alpha\epsilon} = 1 - \alpha\epsilon + \zeta$ where the negligible error term ζ is small compared to ϵ in the sense that $|\zeta|/\epsilon \rightarrow 0$ as $\epsilon \rightarrow 0$. It follows that $\sum_{n \geq 1} p_n(\epsilon) = \alpha\epsilon - \zeta$. If we exclude the possibility of binary stars or multiple stars occupying a given point, i.e., stars really are point masses, then $p_1(\epsilon) = \alpha\epsilon + \zeta_1$ and $\sum_{n \geq 2} p_n(\epsilon) = \zeta_2$ where ζ_1 and ζ_2 are both in absolute value small compared to ϵ . We may hence identify the parameter α as the *density of stars per unit volume*.

Before proceeding further it will be convenient to introduce some order notation to sweep the nuisance, but ultimately negligible, order terms like ζ, ζ_1 , and ζ_2 under an asymptotic carpet. We write $\mathfrak{o}(\epsilon)$ for any term such as ζ, ζ_1, ζ_2 , etc., small in absolute value compared to ϵ in the sense that the ratio of these terms to ϵ tends to zero as $\epsilon \rightarrow 0$.

Now consider the effect of adding a region of infinitesimal volume ϵ to a region of volume v . The combined region will contain a single star if either the original volume v contains precisely one star and the added volume ϵ contains no stars or the volume v contains no stars and the added volume ϵ contains precisely one star. It follows that

$$p_1(v + \epsilon) = p_1(v)p_0(\epsilon) + p_0(v)p_1(\epsilon) = p_1(v)(1 - \alpha\epsilon + \mathfrak{o}(\epsilon)) + e^{-\alpha v}(\alpha\epsilon + \mathfrak{o}(\epsilon)).$$

Rearrange terms, collect all the small-order terms under one umbrella, and divide throughout by ϵ to obtain $\frac{1}{\epsilon}(p_1(v + \epsilon) - p_1(v)) + \alpha p_1(v) = \alpha e^{-\alpha v} + \mathfrak{o}(1)$ where $\mathfrak{o}(1)$ denotes an order term that tends to zero as $\epsilon \rightarrow 0$. Let $\epsilon \rightarrow 0$ on both sides to finally get rid of the pesky order term and obtain the differential equation $p_1'(v) + \alpha p_1(v) = \alpha e^{-\alpha v}$ with the obvious boundary condition $p_1(0) = 0$. Multiplying both sides by $e^{\alpha v}$ simplifies the expression to the form $\frac{d}{dv}(e^{\alpha v}p_1(v)) = \alpha$. With a view to generalisation it is expedient to now introduce some notation.

For each integer $n \geq 0$, define $q_n(v) = e^{\alpha v}p_n(v)$ for $v \geq 0$. Observe the boundary conditions $q_0(0) = p_0(0) = 1$ and $q_n(0) = p_n(0) = 0$ for $n \geq 1$.

With these definitions in hand, we are left with the simple differential equation $q_1'(v) = \alpha$ with boundary condition $q_1(0) = 0$, leading to the general solution $q_1(v) = \alpha v$ for all $v \geq 0$. It follows that $p_1(v) = \alpha v e^{-\alpha v}$.

The general case is not much harder. Again append a region of infinitesimal volume ϵ to a region of volume v . The combined region will contain precisely n stars if volume v contains n and volume ϵ contains none, or volume

v contains $n - 1$ and volume ϵ contains one, or volume v contains $n - k$ and volume ϵ contains k for some $k \geq 2$. The last possibility has probability no larger than $\sum_{k \geq 2} p_k(\epsilon) = o(\epsilon)$. It follows that

$$\begin{aligned} p_n(v + \epsilon) &= p_n(v)p_0(\epsilon) + p_{n-1}(v)p_1(\epsilon) + \sum_{k \geq 2} p_{n-k}(v)p_k(\epsilon) \\ &= p_n(v)(1 - \alpha\epsilon + o(\epsilon)) + p_{n-1}(v)(\alpha\epsilon + o(\epsilon)) + o(\epsilon). \end{aligned}$$

After a rearrangement of terms and simplification we obtain

$$\frac{p_n(v + \epsilon) - p_n(v)}{\epsilon} + \alpha p_n(v) = \alpha p_{n-1}(v) + o(1).$$

Take the limit as $\epsilon \rightarrow 0$ to obtain the recursive differential equation $p'_n(v) + \alpha p_n(v) = \alpha p_{n-1}(v)$. Multiplying both sides by $e^{\alpha v}$ results in the slightly more manageable form $\frac{d}{dv}(e^{\alpha v} p_n(v)) = \alpha(e^{\alpha v} p_{n-1}(v))$ or, in our improvised notation, $q'_n(v) = \alpha q_{n-1}(v)$ with the boundary condition $q_n(0) = 0$ for $n > 0$. Induction now quickly yields the general solution $q_n(v) = (\alpha v)^n / n!$ for $v \geq 0$.

It follows that $p_n(v) = e^{-\alpha v} (\alpha v)^n / n!$ for all $v \geq 0$ and all integers $n \geq 0$. *The number of stars in any region of volume v hence conforms to a Poisson distribution with parameter αv .* ►

The reader will have realised that the geometrical structure of three-dimensional space played no rôle in the analysis. In consequence, the result holds true in any number of dimensions (though the word “volume” is idiosyncratic in one and two dimensions and it is prudent to replace it by “length” and “area”, respectively). We turn to an example in one dimension next.

EXAMPLE 3) *Call arrivals.* The times between successive arrivals of calls at a telephone switchboard may be modelled as an independent sequence of random variables subscribing to a common exponential distribution. What can be said about the distribution of the number of calls $N(t)$ that arrive in an interval of duration t ? Strike out the odd reference to “volume” in the previous example and replace it with “length” or “duration” to conclude that the distribution of calls in time forms a Poisson ensemble in one dimension. It follows that $N(t)$ has a Poisson distribution with parameter αt for some positive α or, more explicitly, $P\{N(t) = k\} = e^{-\alpha t} (\alpha t)^k / k!$ for every integer $k \geq 0$, and an unexpected connection between the exponential and Poisson distributions has emerged. This has important consequences in the theory of queues. ►

7 Waiting times, the Poisson process

Suppose X_1 and X_2 are independent random variables with common exponential density $\alpha e^{-\alpha x}$ for $x > 0$. The sum $S_2 = X_1 + X_2$ then has density

$$g_2(t) = \alpha^2 \int_0^t e^{-\alpha x} e^{-\alpha(t-x)} dx = \alpha(\alpha t)e^{-\alpha t} \quad (t > 0).$$

The general case follows quickly by induction.

THEOREM 1 Suppose X_1, \dots, X_n are independent random variables drawn by independent sampling from the common exponential density $\alpha e^{-\alpha x}$ for $x > 0$. The sum $S_n = X_1 + \dots + X_n$ then has density $g_n(t)$ and d.f. $G_n(t)$ given by

$$g_n(t) = \alpha \frac{(\alpha t)^{n-1}}{(n-1)!} e^{-\alpha t} \text{ and } G_n(t) = 1 - e^{-\alpha t} \sum_{k=0}^{n-1} \frac{(\alpha t)^k}{k!} \quad (t > 0). \quad (7.1)$$

PROOF: The cases $n = 1$ and $n = 2$ having already been covered establish the base of the induction. As induction hypothesis suppose S_n has density $g_n(t)$ given by (7.1) for some positive integer n . As $S_{n+1} = S_n + X_{n+1}$ is the sum of the independent, positive random variables S_n and X_{n+1} , the convolution $g_{n+1} = g_n * g_1$ simplifies to

$$g_{n+1}(t) = \frac{\alpha^{n+1}}{(n-1)!} \int_0^t x^{n-1} e^{-\alpha x} e^{-\alpha(t-x)} dx = \alpha \frac{(\alpha t)^n}{n!} e^{-\alpha t} \quad (t > 0).$$

This completes the induction. The form of the d.f. may now be verified by successive integration by parts. ►

The density g_n is the important gamma density which crops up in a variety of applications. (See Problem 29, for instance.) We will return to the general form of this density in Section 8.

WAITING TIMES

In applications in queuing theory, the exponential distribution is frequently used to model a lack of memory in service times to customers in a queue or to model inter-arrival times for customers (calls, taxis, gondolas) arriving in a queue as we saw in Example 6.3. In these contexts the notion of waiting time or residual time is important.

For definiteness suppose X_1, X_2, \dots represents a sequence of inter-arrival times of customers to a queue. As before, we suppose that the X_i are generated by independent sampling from a common exponential density $\alpha e^{-\alpha x}$ for $x > 0$.

Suppose one observes the system at some “random time” t . The observation point lies in some interval between arrivals, say, $X^{(t)}$ in duration (Figure 11). What can be said about the *residual time* $R^{(t)}$ from the observation point t to the next arrival? This is a typical waiting time problem. Two opposing points of view may now be sketched leading to a disturbing and apparently paradoxical state of affairs.

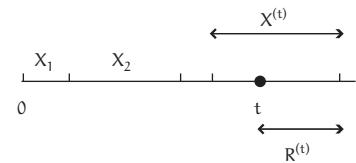


Figure 11: Residual time.

PLAUSIBLE ARGUMENT 1 Given that the k th arrival has not yet transpired by time t , the additional time $R^{(t)}$ that one has to wait before the arrival is an exponential random variable with mean $1/\alpha$ by virtue of the memoryless property of the exponential distribution.

PLAUSIBLE ARGUMENT 2 The random variable $X^{(t)}$ in which the observation point lies is an exponential random variable with mean $1/\alpha$. As the location of the observation point within $X^{(t)}$ is uniform over the interval the residual time $R^{(t)}$ will, on average, have duration $1/(2\alpha)$.

Both points of view cannot simultaneously hold true. This is sometimes called the *inspection paradox* and it caused significant perturbation in the queuing community until it was resolved. As is typical in such cases, the confusion arises because of ambiguity in the problem specification and it is clear where the difficulty lies.

What precisely is meant by the notion of a “random time”? Informally, one might take it to mean a random variable that is independent of the X_i and uniformly distributed over some interval. This, however, is open to the criticism that the length of the interval has not been specified. To add to the confusion, one frequently sees the modifier “random time in the steady state” by which is loosely meant an observation time t after the system has been in operation for an “infinite” amount of time. (After which time, presumably, all transient behaviour has been ironed out and the flow of customers has settled down into a “steady state”.) The manifest absurdity of attempting to model this by a uniform distribution with infinite support leads us to rethink the probabilistic underpinnings of the problem.

As before, write $S_n = X_1 + \dots + X_n$ for each $n \geq 1$; for $n = 0$ it will be convenient and natural to set $S_0 = 0$. Suppose now that the observation instant $t > 0$ is any *fixed* time instant and let $N(t)$ be the largest value of the index k for which $S_k \leq t$. Naïve intuition would suggest that the interval of length $X^{(t)}$ that contains the point t has the common exponential distribution of the rest of the X_i . But does it? Let $F^{(t)}$ and $f^{(t)}$ be the d.f. and density, respectively, of $X^{(t)}$. Then $N(t)$ is the unique index satisfying $S_{N(t)-1} < t \leq S_{N(t)}$ where the notation emphasises the time- and chance-dependent nature of the index. The observation point t hence lies in the interval between the $(N(t) - 1)$ th and $N(t)$ th arrivals so that we identify $X^{(t)} = X_{N(t)}$.

The random index $N(t)$ can range over all integers $k \geq 1$. Thus, for any $x > 0$, the occurrence of the event $\{X^{(t)} \leq x\}$ is equivalent to saying that, for some integer $k \geq 1$, $S_{k-1} < t$ and $t - S_{k-1} < X_k \leq x$. If $k = 1$ this means that the first arrival is after t and before x and so

$$\mathbf{P}\{X_{N(t)} \leq x, N(t) = 1\} = \mathbf{P}\{t < X_1 \leq x\} = \begin{cases} 0 & \text{if } x \leq t, \\ e^{-\alpha t} - e^{-\alpha x} & \text{if } x > t. \end{cases} \quad (7.2)$$

For $k \geq 2$, the random variables S_{k-1} and X_k are independent with marginal gamma densities g_{k-1} and g_1 , respectively, given by (7.1). Integrating over the region \mathbb{A} consisting of the collection of points (s, r) in the plane satisfying the inequalities $0 < s < t$ and $t - s < r < x$, we hence have

$$P\{S_{k-1} < t, t - S_{k-1} < X_k \leq x\} = \iint_{\mathbb{A}} g_{k-1}(s) g_1(r) dr ds.$$

The events $\{S_{k-1} < t, t - S_{k-1} < X_k \leq x\}$ are mutually exclusive and so the contribution to the d.f. of $X^{(t)}$ from the cases $k \geq 2$ is obtained by summing over k . As the form of the gamma density (7.1) shows that $\sum_{k=2}^{\infty} g_{k-1}(s) = \alpha$ for every $s > 0$, we obtain

$$\sum_{k=2}^{\infty} \iint_{\mathbb{A}} g_{k-1}(s) g_1(r) dr ds = \iint_{\mathbb{A}} \sum_{k=2}^{\infty} g_{k-1}(s) g_1(r) dr ds = \iint_{\mathbb{A}} \alpha g_1(r) dr ds,$$

the interchange in the order of summation and integration occasioning no comment as the integrands are positive and the series convergent. The expression on the right represents the probability that $X_{N(t)} \leq x$ and $N(t) \geq 2$. To evaluate the integral over the region \mathbb{A} we observe that the set of points \mathbb{A} in the plane corresponds to the shaded regions shown in Figure 12, the two distinct cases

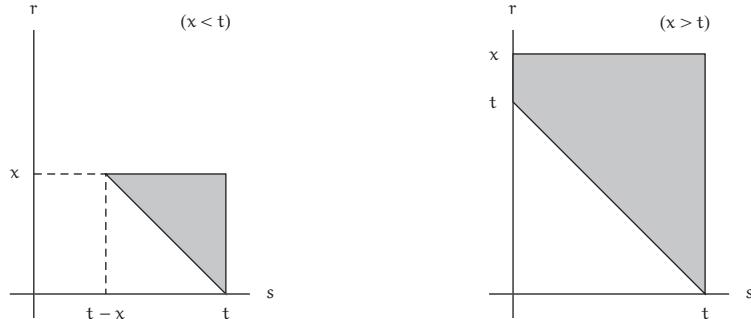


Figure 12: The interval containing the observation point t .

manifest depending on whether $x \leq t$ or $x > t$. We may consolidate them by observing that, in either case, s varies in the interval $\max\{0, t - x\} < s \leq t$ and, for each s in this interval, r varies in the range $t - s < r \leq x$. By evaluating the iterated integral we hence obtain

$$\begin{aligned} P\{X_{N(t)} \leq x, N(t) \geq 2\} &= \iint_{\mathbb{A}} \alpha g_1(r) dr ds = \alpha \int_{\max\{0, t-x\}}^t \int_{t-s}^x g_1(r) dr ds \\ &= \alpha \int_{\max\{0, t-x\}}^t (e^{-\alpha(t-s)} - e^{-\alpha x}) ds = \begin{cases} 1 - e^{-\alpha x} - \alpha x e^{-\alpha x} & \text{if } x \leq t, \\ 1 - e^{-\alpha t} - \alpha t e^{-\alpha x} & \text{if } x > t. \end{cases} \quad (7.3) \end{aligned}$$

Combining the expressions from (7.2,7.3), we obtain the distribution

$$F^{(t)}(x) = P\{X^{(t)} \leq x\} = 1 - e^{-\alpha x} - \alpha \min\{x, t\}e^{-\alpha x} \quad (x > 0).$$

A separate differentiation of the d.f. in the two cases $x \leq t$ and $x > t$ now yields the density.

THEOREM 2 *For every $t > 0$, the density of $X^{(t)}$ is given by*

$$f^{(t)}(x) = \begin{cases} \alpha^2 x e^{-\alpha x} & \text{if } 0 < x \leq t, \\ \alpha(1 + \alpha t)e^{-\alpha x} & \text{if } x > t. \end{cases}$$

The specific form of the density is not of particular significance. The key point here is that $X^{(t)}$ does not have the common exponential density of the X_i ; the density indeed depends on the observation point t . The notion of a “steady state” may be accommodated in this framework by allowing t to tend to infinity and we observe that the density of $X^{(t)}$ tends to the limiting gamma density $\alpha^2 x e^{-\alpha x}$ ($x > 0$) as $t \rightarrow \infty$. The limiting density is not exponential either.

One may explain this paradoxical result by considering a slightly different setting. Suppose a *random* observation point t is chosen by sampling from a uniform distribution over a large range $(0, \tau)$. If there are m arrivals in $(0, \tau)$ then t may lie in any of the $m+1$ inter-arrival intervals but *it is not equally likely that it will be in any of them*. Indeed, the larger intervals occupy a disproportionate portion of the real estate in $(0, \tau)$ and so t is much more likely to lie in such intervals. This leads to a density with a heavier tail than initially anticipated.

A similar analysis yields the density of the residual time $R^{(t)} = S_{N(t)} - t$. The event $\{R^{(t)} > x\}$ occurs if, and only if, for some integer $k \geq 1$, we have $S_{k-1} \leq t$ and $X_k > t - S_{k-1} + x$. Proceeding as before, we obtain

$$P\{R^{(t)} > x\} = e^{-\alpha(t+x)} + \alpha \int_0^t \int_{t-s+x}^{\infty} g_1(r) dr ds = e^{-\alpha x} \quad (x > 0),$$

the result independent of the selection of observation point t .

THEOREM 3 *The residual time $R^{(t)}$ has the exponential density with mean $1/\alpha$.*

This result is in accordance with an unthinking use of the memoryless property but, in view of Theorem 2, the reader may well feel that Theorem 3 has an accidental character. Certainly there is a subtlety in the analysis that naïve intuition did not prepare us for. The general setting leads to the *Pollaczek–Khinchin formulæ* covered in Section XIV.5.

THE POISSON PROCESS, REVISITED

We recall that with the X_i representing inter-arrival times of customers, the partial sum $S_n = X_1 + \dots + X_n$ represents the time of the n th arrival. In this setting,

for each positive t , the random index $N(t) = \sup\{n : S_n \leq t\}$ has evidentiary value in representing the number of arrivals up till time t . The fact that the partial sums S_n are governed by gamma distributions provides a direct path to analysing the behaviour of the number of arrivals.

The event $\{N(t) = n\}$ occurs if, and only if, $S_n \leq t$ and $S_{n+1} > t$. As the event $\{S_{n+1} > t\}$ is comprised of the disjoint union of the events $\{S_n > t\}$ and $\{N(t) = n\}$, by additivity

$$P\{S_{n+1} > t\} = P\{S_n > t\} + P\{N(t) = n\}.$$

By Theorem 1, it follows that

$$\begin{aligned} P\{N(t) = n\} &= P\{S_{n+1} > t\} - P\{S_n > t\} = [1 - G_{n+1}(t)] - [1 - G_n(t)] \\ &= e^{-\alpha t} \sum_{k=0}^n \frac{(\alpha t)^k}{k!} - e^{-\alpha t} \sum_{k=0}^{n-1} \frac{(\alpha t)^k}{k!} = e^{-\alpha t} \frac{(\alpha t)^n}{n!} \quad (n \geq 0; t > 0). \end{aligned}$$

Thus, for each positive t , the random variable $N(t)$ has a Poisson distribution with mean αt as we saw by another route in Example 6.3.

Somewhat more can be said. In a slight abuse of notation, write $N(a, b] = N(b) - N(a)$ for the number of arrivals in the interval $(a, b]$ where $0 < a < b$. We call $N(a, b]$ an *increment*. The observation point $t = a$ resets the clock and the waiting time till the first arrival after this point is governed by the residual time distribution which, in view of Theorem 3, shares the common exponential distribution of the X_i and is independent of the other members of the sequence. It follows that $N(a, b]$ is a probabilistic replica of $N(t) = N(0, t]$ with $t = b - a$ and is hence governed by the Poisson distribution with mean $\alpha(b - a)$. Furthermore, conditioned on m arrivals up till time a , in notation, $N(0, a] = m$, the increment $N(a, b]$ is determined by the random sequence $R^{(a)} = X'_{m+1}, X_{m+2}, X_{m+3}, \dots$ which is independent of X_1, \dots, X_m and is a probabilistic replica of the original exponential sequence $\{X_i, i \geq 1\}$. It follows that $N(0, a]$ and $N(a, b]$ are independent increments. By induction, for every k and every sequence of time instants $0 = t_0 < t_1 < t_2 < \dots < t_k$, the random variables $N(t_{j-1}, t_j] = N(t_j) - N(t_{j-1})$ ($1 \leq j \leq k$) are independent. This pithy observation deserves recording as a slogan.

SLOGAN *The Poisson arrival process has independent increments.*

Viewed as a function of time, $\{N(t), t \geq 0\}$ determines a random process whose sample paths induced by the sample points (X_1, X_2, \dots) have a step character as illustrated in Figure 13. This is the celebrated *Poisson process*.

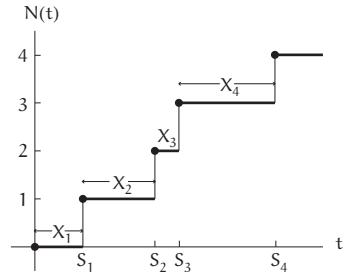


Figure 13: A sample path of a Poisson process.

8 Densities arising in queuing theory

Considerations of the uniform and the exponential densities engender other related densities. These occur occasionally in applications.

The gamma density made its appearance as the convolution of exponential densities in Section 7. In this setting it appeared indexed by an integral parameter k in the generic form $g_k(x) = \alpha \frac{(\alpha x)^{k-1}}{(k-1)!} e^{-\alpha x}$ ($x > 0$). The form of the density may be extended to non-integral, positive parameters as we now see. Let $\alpha > 0$ and $\nu > 0$ be fixed, but arbitrary, real numbers. The *gamma density with parameters α and ν* is defined by

$$g_\nu(x; \alpha) = \frac{\alpha}{\Gamma(\nu)} (\alpha x)^{\nu-1} e^{-\alpha x} \quad (x > 0); \quad (8.1)$$

if the parameter α is clear from the context we simplify notation and write simply $g_\nu(x) = g_\nu(x; \alpha)$. The normalising factor in the denominator is just the gamma function (hence the name of the density) which we recall is defined by

$$\Gamma(\nu) = \int_0^\infty t^{\nu-1} e^{-t} dt \quad (\nu > 0).$$

As $\Gamma(k) = (k-1)!$ for any integer $k \geq 1$, the definition reduces to the case seen in (7.1) if ν is an integral parameter. The reader may verify that the mean of the gamma density $g_\nu(x; \alpha)$ is ν/α while its variance is ν/α^2 . (The simplest path is to massage the integrand into the form of a gamma density; simple manipulations show that the mean is then given by $\mu = \Gamma(\nu+1)/\alpha\Gamma(\nu)$ while the variance is given by $\sigma^2 = \Gamma(\nu+2)/\alpha^2\Gamma(\nu) - \mu^2$. The characteristic gamma function identity $\Gamma(\nu+1) = \nu\Gamma(\nu)$ completes the job.)

The constant α is the trivial scale parameter. To focus on the essential role of the parameter ν , set $\alpha = 1$. With $x > 0$, the densities $g_1(x; 1) = e^{-x}$, $g_{1/2}(x; 1) = e^{-x}/\sqrt{\pi x}$, and $g_2(x; 1) = xe^{-x}$ illustrate the three general behaviours of the family (see Figure 14). If $\nu = 1$ we obtain the standard exponential density; of course, this density decreases monotonically with x in the positive half-axis. When $\nu < 1$ the density again decreases monotonically over the positive half-axis but is now unbounded near the origin. Finally, for $\nu > 1$ we obtain a unimodal density that attains its maximum at the point $x = \nu - 1$.

EXAMPLE: Waiting time in a single-server queue. In a packet-switched communication network, data packets arriving at a node, for instance, a transmission line, are enqueued and wait their turn for transmission. In a standard model packet lengths are assumed to be independent random variables with a common exponential distribution with mean $1/\alpha$. If an arriving packet finds k packets ahead of it (including the one in transmission), what is the waiting time before its transmission can commence? The memoryless property of the exponential ensures that the packet currently under transmission still has a residual

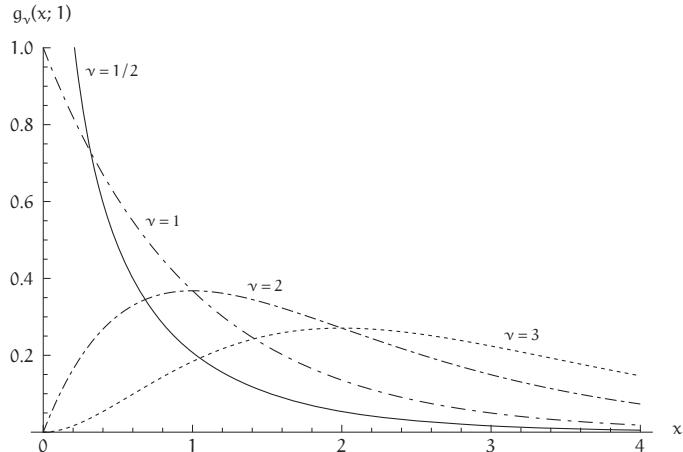


Figure 14: Gamma densities.

length that is exponential with mean $1/\alpha$ (see Theorem 7.3). It follows that the packet waiting time for transmission is the sum of k independent, exponential random variables, hence has the gamma density $g_k(x; \alpha)$. The total sojourn time of the packet (including its transmission time) is the sum of $k + 1$ exponential random variables, hence has the gamma density $g_{k+1}(x; \alpha)$. An extension to the n -server queue is considered in Problem 36.

The limitations of the model are clear as it ignores granularity at the level of the packets which are comprised of bits. In consequence, for small typical packet sizes the model provides optimistic results. When packet sizes are large the granularity effects are not nearly so pronounced. ▶

A variety of distributions important in statistics are derived from the gamma density and we consider some of these in Section X.2. The gamma density also appears front and centre in queuing theory in which context it is called *Erlangian* after the Danish pioneer A. K. Erlang who established much of the basic theory.

The gamma density is the continuous intellectual cousin of the negative binomial distribution and inherits the same closure property.

THEOREM *For each fixed $\alpha > 0$, the family of gamma densities $\{g_\nu(\cdot; \alpha), \nu > 0\}$ is closed under convolutions, that is, $(g_\mu * g_\nu)(x) = g_{\mu+\nu}(x)$ for every μ and ν .*

PROOF: As the gamma densities have support in the positive half-axis, for any $x > 0$ we have

$$(g_\mu * g_\nu)(x) = \int_0^x g_\mu(t)g_\nu(x-t) dt = \frac{\alpha^{\mu+\nu} e^{-\alpha x}}{\Gamma(\mu)\Gamma(\nu)} \int_0^x t^{\mu-1} (x-t)^{\nu-1} dt.$$

The change of variable of integration $xs \leftarrow t$ results in the simplified expression

$$(g_\mu * g_\nu)(x) = \frac{\alpha^{\mu+\nu} x^{\mu+\nu-1} e^{-\alpha x}}{\Gamma(\mu)\Gamma(\nu)} \int_0^1 s^{\mu-1} (1-s)^{\nu-1} dt.$$

The expression $B(\mu, \nu) = \int_0^1 s^{\mu-1} (1-s)^{\nu-1} ds$ is called the *beta function*. What is germane here is that $B(\mu, \nu)$ is a positive constant. We hence have $(g_{\alpha,\mu} * g_{\alpha,\nu})(x) = C\alpha(\alpha x)^{\mu+\nu-1} e^{-\alpha x}$ for a positive constant C . It only remains to determine the value of this normalising constant. But $g_\mu * g_\nu$ is a density (why?) and a comparison with (8.1) shows that C must be equal to $1/\Gamma(\mu+\nu)$. It follows that $g_\mu * g_\nu$ is indeed the gamma density $g_{\mu+\nu}$. ▶

That the result is not unexpected may be seen from the following argument. Suppose k and l are positive integers. Suppose X_1, \dots, X_{k+l} are independent random variables with a common exponential distribution with mean $1/\alpha$. Then via Theorem 7.1, the sum $S_1 = X_1 + \dots + X_k$ has the gamma density $g_k(x) = g_k(x; \alpha)$, where we again suppress the dependence on the fixed parameter α ; likewise, the sum $S_2 = X_{k+1} + \dots + X_{k+l}$ has the gamma density $g_l(x)$. As S_1 and S_2 are independent, it follows that the density of $S_1 + S_2$ is given by the convolution $(g_k * g_l)(x)$. On the other hand, $S_1 + S_2 = X_1 + \dots + X_{k+l}$ so that we identify this density with the gamma density $g_{k+l}(x)$. What is new is that this closure property of the gamma densities holds, for fixed α , even when the parameter ν is allowed to vary in a non-integral range.

9 Densities arising in fluctuation theory

As a side product of the proof of the theorem of the previous section, we've also shown perforce that the beta function is given by

$$B(\mu, \nu) = \int_0^1 s^{\mu-1} (1-s)^{\nu-1} ds = \frac{\Gamma(\mu)\Gamma(\nu)}{\Gamma(\mu+\nu)}.$$

This normalisation suggests a new family of densities indexed by real numbers $\mu > 0$ and $\nu > 0$. The *beta density with parameters μ and ν* is defined by

$$\beta_{\mu,\nu}(x) = \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} x^{\mu-1} (1-x)^{\nu-1} \quad (0 < x < 1).$$

When $\mu = \nu = 1$ we obtain $\beta_{1,1}(x) = 1$ in the unit interval so that the uniform density is contained within the family of beta densities. (Analogies with the exponential and the gamma go only so far, however. The beta densities are not closed under convolution (manifestly) and convolutions of uniform densities do not yield beta densities.) When $\mu < 1$ and $\nu < 1$ the graph of $\beta_{\mu,\nu}(x)$ is concave upwards in the unit interval with a unique interior minimum and

unbounded at the endpoints $x = 0$ and $x = 1$. When $\mu > 1$ and $\nu > 1$ the graph of $\beta_{\mu,\nu}(x)$ is bell-shaped with a unique interior maximum while taking a value of zero at the endpoints $x = 0$ and $x = 1$ as we see in Figure 15.

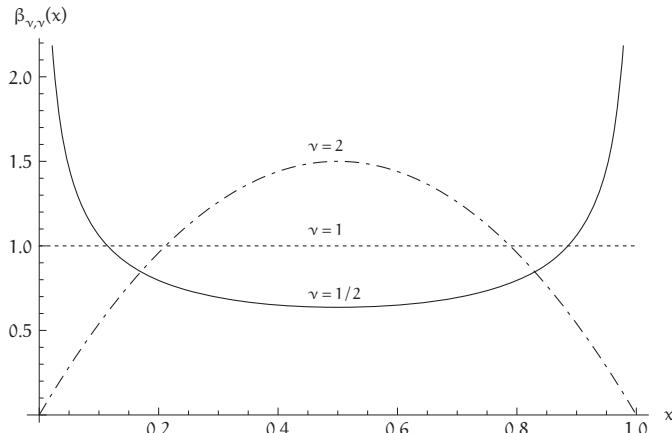


Figure 15: Beta densities with $\mu = \nu$.

The parameter μ in the density is not to be confused with the mean. Indeed, trite manipulations show that the mean of the density $\beta_{\mu,\nu}$ is $\mu/(\mu+\nu)$ while its variance is $\mu\nu/(\mu+\nu)^2(\mu+\nu+1)$ as the reader may verify. (Again, relate the integrands to the form of another beta density.)

EXAMPLE: Uniform and beta-distributed priors. The fact that the uniform density is contained in the beta family explains the occasional appearance of the beta density in Bayesian analysis. While it is not our intent to get embroiled in a dispute between Bayesians and non-Bayesians, the setting particularised to coin-tossing experiments is briefly as follows. In situations involving binomial experiments with an unknown success probability, the Bayesian acolyte models the success probability as arising from a prior random experiment according to a given distribution (the “prior”). If nothing whatsoever is known about the success probability in the experiment it is tempting, following Laplace, to assume that the success probability is uniformly distributed in the unit interval. Thus, the Bayesian imagines that nature presents us with a coin chosen at random. This policy at least has the advantage of “assuming the least” about the putative underlying distribution.⁷

Adopting a Bayesian outlook, if, however, one feels that the uniform prior is too restrictive an assumption for the problem at hand then one may

⁷This is the basic philosophical principle underlying the “maximum entropy” formulation popularised by E. T. Jaynes.

elect to replace it speculatively with a beta density as a prior on the somewhat specious rationale that the family of beta densities not only has the virtue of including the uniform as a member but now provides a much richer set of alternatives that one hopes is better attuned to the problem at hand.

There will be little dispute about such a procedure if there is strong physical intuition backed by data validating the choice of prior. Tacit in such a statement is the assumption that a large number of such experiments have been performed so that reliable statistical data are available. Absent convincing physical intuition there are philosophical problems with the assignment of priors if there is not much data or if what there is is inconclusive. In this case the ritualistic assignment of probabilities to priors is logically barren as one crosses the boundary between a statistically rigorous procedure and a metaphysical, faith-based argument. For a cautionary example, the reader is referred to the analysis of Laplace's law of succession (Section II.8). As W. Feller has remarked, even strong faith must be fortified by a statistical test. ►

The beta density with $\mu = \nu = 1/2$ crops up unexpectedly in the theory of fluctuations. This is the famous but inappropriately named *arc sine density*

$$\beta_{\frac{1}{2}, \frac{1}{2}}(x) = \frac{1}{\pi\sqrt{x(1-x)}} \quad (0 < x < 1)$$

with mean 1/2 and variance 1/8 whose graph appears in Figure 15. The origins of the name appear obscure until one considers the corresponding distribution function

$$\int_{-\infty}^x \beta_{\frac{1}{2}, \frac{1}{2}}(t) dt = \int_0^x \frac{1}{\pi\sqrt{t(1-t)}} dt = \frac{2}{\pi} \int_0^{\sqrt{x}} \frac{du}{\sqrt{1-u^2}} = \frac{2}{\pi} \arcsin\sqrt{x}$$

(the change of variable to $u^2 \leftarrow t$ inside the integral is perspicuous); hence the name. A series of investigations beginning with P. Lévy showed the underlying connections between the mysterious appearance of the arc sine density in various applications and the theory of fluctuations in random walks. Several distinguished names including K. L. Chung, P. Erdős, W. Feller, M. Kac, and E. Sparre Andersen have since fleshed out much of the basic theory. A key connection was pointed out in the arc sine law for last visits in Section VIII.5.

10 Heavy-tailed densities, self-similarity

Suppose X is a beta-distributed random variable with parameters μ and ν . The transformation $Y = 1/X$ then creates a new random variable with support in $(1, \infty)$ whose distribution function is given by

$$P\{Y \leq y\} = P\{X \geq 1/y\} = \int_{y^{-1}}^1 \beta_{\mu, \nu}(x) dx \quad (y > 1)$$

and differentiation shows that Y has density

$$\frac{\beta_{\mu,\nu}(y^{-1})}{y^2} = \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} \frac{(y-1)^{\nu-1}}{y^{\mu+\nu}} \quad (y > 1).$$

It is easy to see that this density asymptotically satisfies the power law $y^{-\mu-1}$ which decays rather slower than, for instance, the exponential decay seen in the exponential and gamma densities. Such power law densities are accordingly said to have *heavy tails*.

Densities with an asymptotic power law dependence of this form are called *Pareto* after the Italian economist Vilfredo Pareto who used it to model income distribution in the late nineteenth century to account for the observed fact that a relatively small proportion of the population controlled a significant portion of the resources. (References to the Pareto principle, 80% of the consequences stem from 20% of the causes—or, there are few billionaires but they control much of the flow of money—may still be seen in the literature.)

Cast in a pure power law form, for every $\mu > 0$ we may define the Pareto density with support in $(1, \infty)$ by $p_\mu(x) = \mu x^{-\mu-1}$. If the random variable I representing the income of a random individual is postulated to have density p_μ then, with a minimum income of one unit, $P\{I > x\} = x^{-\mu}$. Plotted on a log-log graph, the tail of the Pareto distribution hence exhibits a characteristic constant slope. This has the interesting consequence that $P\{I > a^2 \mid I > a\} = a^{-2\mu}/a^{-\mu} = a^{-\mu}$ and, in general, $P\{I > a^{n+1} \mid I > a^n\} = a^{-\mu}$ for all non-negative integers n . Thus, the distribution of mass is unaltered in every subinterval. If 80% of the wealth is concentrated in the top 20% of incomes then 80% of the wealth in the top 20% is concentrated in the top 20% of that group, and so on. This property is called *self-similarity*.

Trivial changes in scale give rise to other distributions of this type: for every positive a , the random variable $I_a = aI$ has the Pareto density $p_{a,\mu}(x) = \frac{1}{a} p_\mu\left(\frac{x}{a}\right) = \mu a^\mu x^{-\mu-1}$ with support in (a, ∞) and a right tail $P\{I_a > x\} = \left(\frac{a}{x}\right)^\mu$ for $x \geq a$.

In addition to Pareto's original work on income distribution, power law densities of this form have been used to model a variety of phenomena including city sizes, stock indices, word frequencies (in which context the distributions are called *Zipf* after the Harvard linguist George Kingsley Zipf), machine failures, and more recently, internet traffic on account of observations that suggest that the statistical fluctuations in internet traffic appear to be replicated on any time scale, in other words, that the traffic appears to have a self-similar characteristic.

It is trivial to check that for $\mu > 1$ the Pareto density p_μ has mean $\mu/(\mu-1)$ but that for $0 < \mu \leq 1$ the mean is $+\infty$. The critical exponent is $\mu = 1$ in which case the density decays quadratically. A two-sided relative of the Pareto density with quadratic decay may be formulated as $c(x) = \pi^{-1}(1+x^2)^{-1}$. This is the standard *Cauchy density*. A routine integration may be performed to show

that the corresponding distribution function is

$$C(x) = \int_{-\infty}^x c(t) dt = \frac{1}{\pi} \arctan t \Big|_{-\infty}^x = \frac{1}{\pi} \left(\arctan x + \frac{\pi}{2} \right).$$

The integral $\int_{-\infty}^{\infty} \frac{x}{1+x^2} dx$ does not converge absolutely and the Cauchy density hence does not have a mean.

Scaling by the positive a results in the generic form of the Cauchy density $c_a(x) = \frac{1}{a} c(\frac{x}{a}) = a/(\pi(a^2 + x^2))$ with the corresponding distribution function $C_a(x) = \frac{1}{\pi} (\arctan \frac{x}{a} + \frac{\pi}{2})$. The Cauchy densities exhibit a remarkable stability property.

THEOREM (STABILITY OF THE CAUCHY DENSITY) *The convolution relation $c_a * c_b = c_{a+b}$ holds for every choice of positive parameters a and b .*

PROOF: The convolution integral $(c_a * c_b)(x)$ may be evaluated by a routine, if tedious, exercise in partial fractions. If one is willing to delve into complex numbers then Fourier transforms offer an elegant alternative and I will sketch the argument here leaving the direct calculation to the motivated reader.

To each integrable function $f(x)$ we may associate the Fourier transform $\hat{f}(\zeta) = \int_{-\infty}^{\infty} f(x) e^{i\zeta x} dx$. Recall that Fourier theory tells us that f is also completely specified by \hat{f} . The convolution property of Fourier transforms tells us that the Fourier transform of the convolution $(c_a * c_b)(x)$ is given by $\widehat{c_a * c_b}(\zeta) = \widehat{c_a}(\zeta) \widehat{c_b}(\zeta)$ so that it will suffice, in principle, to evaluate the Fourier transform of the generic Cauchy density. Now

$$\widehat{c_a}(\zeta) = \frac{a}{\pi} \int_{-\infty}^{\infty} \frac{e^{i\zeta x}}{a^2 + x^2} dx.$$

The simplest approach to evaluating the integral is via Cauchy's residue theorem which quickly yields $\widehat{c_a}(\zeta) = e^{-a|\zeta|}$. It follows that $\widehat{c_a * c_b}(\zeta) = \widehat{c_a}(\zeta) \widehat{c_b}(\zeta) = e^{-(a+b)|\zeta|}$. But the right-hand side is exactly the Fourier transform of $c_{a+b}(x)$ and the uniqueness of the Fourier transform forces the conclusion $c_a * c_b = c_{a+b}$. ▶

As the parameter a in the density $c_a(x)$ corresponds to a positive scale factor, the result asserts that for the Cauchy density the *type* is *stable* under convolutions. While this looks superficially like the corresponding convolution property for the gamma densities seen in Section 8, there is a critical difference. For the preservation of the gamma density under convolution it is essential that the scale parameter α be held fixed while the parameter ν varies; in particular, for the gamma density, the type is *not* stable under convolution.

The Cauchy density is not unique in its stability; as we shall see shortly, the normal density also shares type stability with the Cauchy and other stable densities can also be exhibited.

EXAMPLE: Huygens's principle. Consider a light source situated at the origin of the (x, y) plane emitting a ray at a random angle in the upper half-plane. By

random we mean that the angle Θ made by the ray with the y -axis is uniformly distributed in the interval $(-\pi/2, \pi/2)$. For any given $a > 0$, the ray intersects the line $y = a$ parallel to the axis at a random point $X = a \tan \Theta$. The d.f. of X is then given by $P\{X \leq x\} = P\{-\pi/2 < \Theta \leq \arctan x/a\} = C_a(x)$. It follows that the light intensity along the line $y = a$ has the Cauchy density $c_a(x)$.

In his influential work, *Traité de la Lumière*, Hans Christian Huygens articulated the basic idea that each point on a propagating wavefront of light may be thought of as originating a secondary spherical point source with the wavefront at a subsequent instant formed as the envelope of the wavefronts from the secondary sources. Returning to our example, the density of light intensity along the line $y = a$ is $c_a(x)$ and, likewise, the density along the line $y = a + b$ is $c_{a+b}(x)$; see Figure 16. But, by Huygens's principle, the density of light intensity along the line $y = a + b$ can be recovered by populating the line $y = a$ with secondary sources with density $c_a(x)$; the principle decrees then that the intensity along the line $y = a + b$ must be governed by $(c_a * c_b)(x)$. This leads to the conclusion that $c_{a+b}(x) = (c_a * c_b)(x)$ as the stability of the Cauchy density also asserts. ▶

Suppose X_1, X_2, \dots are independent random variables with each X_i drawn from a Cauchy density $c_{a_i}(x)$. By induction it follows that the sum $S_n = X_1 + \dots + X_n$ has the Cauchy density $c_{a_1+\dots+a_n}(x)$. In particular, if the independent X_i all share the same Cauchy density $c_a(x)$ then the normalised sum $\frac{1}{n} S_n$ also has the density $c_a(x)$ common to the X_i . At first blush this may be vaguely troubling as a “law of averages” appears to be violated. If X_1, \dots, X_n denote the positions of the ray along the line $y = a$ in n independent performances of the experiment then $\frac{1}{n} S_n$ is the position of the ray averaged over the n experiments. Untrained intuition does not prepare us for the fact that the average, far from being near zero as one might anticipate, is actually just as random as the result of a single experiment.

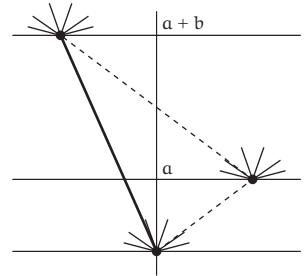


Figure 16: Primary and secondary sources.

11 Problems

1. *Intersecting chords.* Four points X_1, X_2, X_3, X_4 are chosen independently and at random on a circle (say of unit circumference for definiteness). Find the probability that the chords $X_1 X_2$ and $X_3 X_4$ intersect.
2. Find the probability that the quadratic $\lambda^2 - 2U\lambda + V$ has complex roots if the coefficients U and V are independent random variables with a common uniform density $f(x) = 1/a$ for $0 < x < a$.

3. Suppose X_1, X_2, \dots is a sequence of independent random variables uniformly distributed in the unit interval $[0, 1]$. Let $0 < t < 1$ be fixed but arbitrary and let N denote the smallest integer n such that $X_1 + \dots + X_n > t$. Show that $P\{N > n\} = t^n/n!$ and thence find the mean and variance of N .

4. Suppose X is a random point on the unit circle in \mathbb{R}^2 . Determine the density of the length of its projection along the x -axis and find its mean.

5. *Interval lengths for random partitions.* Suppose X_1, \dots, X_n are independent and uniformly distributed over the interval $[0, t)$. These points partition the interval into $n + 1$ subintervals whose lengths, taken in order from left to right, are L_1, L_2, \dots, L_{n+1} . Denote by $p_n(s; t) = P\{\min L_k > s\}$ the probability that all $n + 1$ intervals are longer than s . Prove the recurrence relation

$$p_n(s; t) = \frac{n}{t^n} \int_0^{t-s} x^{n-1} p_{n-1}(s; x) dx,$$

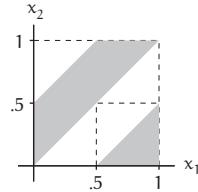
and conclude that $p_n(s; t) = t^{-n} (t - (n + 1)s)_+^n$.

6. An isosceles triangle is formed by a unit vector in the x -direction and another in a random direction. Determine the distribution of the length of the third side in \mathbb{R}^2 and in \mathbb{R}^3 .

7. A ray enters at the north pole of a unit circle, its angle with the positive x -axis being uniformly distributed over $(-\pi/2, \pi/2)$. Find the distribution of the length of the chord within the circle.

8. A stick of unit-length is broken in n (random) places. Show that the probability that the $n + 1$ pieces can form a polygon is $1 - (n + 1)2^{-n}$.

9. *An example of Robbins.* The pair of random variables (X_1, X_2) is uniformly distributed in the shaded region of area $1/2$ inside the unit square as shown in the figure alongside. Let $Y = X_1 + X_2$. (The following questions can be answered geometrically *without any calculation* by referring to the figure. Of course, direct calculation works as well, albeit with a little more effort.) Determine the marginal densities of X_1 , X_2 , and Y . What are the mean and variance of Y ? [Careful! Are X_1 and X_2 independent? Compare with (1.1).]



10. *Closure under convolution of the Cauchy density.* Suppose X and Y are independent Cauchy random variables with density $c_a(x) = a/\pi(a^2 + x^2)$. Show that the variables $X + Y$ and $X + X$ have the same density.

11. *Buffon's needle on a grid.* Consider the plane scored by a grid of parallel lines, a units apart in the x -direction and b units apart in the y -direction. If M. Buffon drops his needle of length r ($< \min\{a, b\}$) at random on the grid, show that the probability that the needle intersects a line equals $r(2a + 2b - r)/(\pi ab)$.

12. *Buffon's cross.* A cross comprised of two unit-length orthogonal segments welded together at their midpoints is dropped on the plane scored by parallel lines one unit apart. Let Z be the number of intersections of the cross with the parallel lines. Show that $E(Z/2) = 2/\pi$ and $\text{Var}(Z/2) = \frac{1}{\pi}(3 - \sqrt{2}) - \frac{4}{\pi^2}$. Comment on whether it is better to estimate π using the needle or the cross.

13. Record values. Denote by X_0 a speculator's financial loss at a downturn in the stock market. Suppose that his friends suffer losses X_1, X_2, \dots . We suppose that X_0, X_1, \dots are independent random variables with a common distribution. The nature of the distribution does not matter but as the exponential distribution serves as a model for randomness, we assume that the X_j are exponentially distributed with common density $f(x) = \alpha e^{-\alpha x}$ for $x > 0$. Here α is positive and α^{-1} is hence the mean of the X_j . To find a measure of our rash gambler's ill luck we ask how long it will be before a friend experiences worse luck: the *waiting time* N is the value of the smallest subscript n such that $X_n > X_0$. Determine the distribution of N , that is to say, find $p(n) := P\{N = n\}$ for each $n = 1, 2, \dots$. What can we say about the expected waiting time before we find a friend with worse luck? [The direct approach by conditioning requires a successive integration of an n -dimensional integral. It is not difficult but you will need to be careful about the limits of integration. There is also a simple direct path to the answer.]

14. Continuation. With X_0 and X_1 as in the previous problem, let $R = X_0/X_1$. Find the distribution function $F(r) := P\{R \leq r\}$ of R . Determine hence the probability that $X_0 > 3X_1$. Does the gambler have cause to complain of ill luck? What can we say about the expected value of R ?

15. Suppose X, Y , and Z are independent with a common exponential distribution. Determine the density of $(Y - X, Z - X)$.

16. Suppose X, Y , and Z are independent, exponentially distributed random variables with means λ^{-1}, μ^{-1} , and ν^{-1} , respectively. Determine $P\{X < Y < Z\}$.

17. Pure birth processes. Starting from an initial state E_0 , a system passes through a sequence of states $E_0 \mapsto E_1 \mapsto \dots \mapsto E_n \mapsto \dots$, staying at E_j for a sojourn time X_j governed by the exponential distribution with mean α_j^{-1} . Then $S_n = X_0 + X_1 + \dots + X_n$ is the epoch at which there is a transition $E_n \mapsto E_{n+1}$. Let $P_n(t)$ denote the probability that at time t the system is in state E_n . Show that $P'_0(t) = -\alpha_0 P_0(t)$ and, for $n \geq 1$, the process satisfies the differential equation $P'_n(t) = -\alpha_n P_n(t) + \alpha_{n-1} P_{n-1}(t)$.

18. The absent-minded professor and the tardy student. A professor (absent-minded by convention) gives each of two students an appointment at 12 noon. One of the students arrives on time; the second arrives 5 minutes late. The amount of time the first student spends closeted with the professor is X_1 ; the second student engages the professor as soon as he is free and spends time X_2 in conference with him. Suppose X_1 and X_2 are independent random variables with the common exponential distribution of mean 30 minutes. What is the expected duration from the arrival of the first student to the departure of the second? [Condition on X_1 .]

19. A waiting time distribution. Suppose X and T are independent random variables, X exponentially distributed with mean $1/\alpha$ and T exponentially distributed with mean $1/\beta$. Determine the density of $X - Y$.

20. Continuation. A customer arriving at time $t = 0$ occupies a server for an amount of time X . A second customer arriving at time $T > 0$ waits for service if $T < X$ and otherwise receives attention immediately. Her waiting time is hence $W = (X - T)^+$. With X and T as in the previous problem, determine the density of W and its mean.

21. Spacings via the exponential distribution. Suppose X_1, \dots, X_n are independent variables with a common exponential distribution and let $S_n = X_1 + \dots + X_n$. Consider the variables $U_j = X_j/S_n$ for $1 \leq j \leq n - 1$ and set $U_n = S_n$. Show that (U_1, \dots, U_n)

has the same distribution as if U_j were the j th spacing in a random partition of the unit interval by $n - 1$ points. [Hint: Evaluate the Jacobian of the transformation.]

22. Suppose X and Y are independent variables with gamma densities $g_m(\cdot; \alpha)$ and $g_n(\cdot; \alpha)$, respectively. Find the joint density of the variables $R = X/(X + Y)$ and $S = X + Y$ and deduce that they are independent.

23. *Continuation.* Suppose X_0, X_1, \dots, X_n are independent variables where, for each j , X_j has gamma density $g_{m_j}(\cdot; \alpha)$. For $1 \leq j \leq n$, define $R_j = X_j/(X_0 + \dots + X_j)$. Show that R_1, \dots, R_n are independent.

24. *Randomisation of the Poisson process.* Suppose $N(t) = N(t; \alpha)$ represents the Poisson arrival process of rate α of Section 7 and let P_n denote the probability that there are exactly n arrivals in the interval $(0, t]$. (a) *Randomised time.* Suppose that, with α fixed, the time parameter t is obtained by sampling from the exponential distribution with mean $1/\beta$. By randomising with respect to t show that P_n has the form of a geometric distribution. (b) *Stratification, randomised rate.* Suppose that, with t fixed, the rate parameter α is obtained by sampling from the gamma density $g_m(\cdot; \beta)$. This models, for instance, arrival rates in different locales. By randomising with respect to α , show now that $P_n = \binom{-m}{n} \left(\frac{-t}{\beta+t}\right)^n \left(\frac{\beta}{\beta+t}\right)^m$ has the form of a negative binomial distribution.

25. *Dirichlet density.* With X_0, X_1, \dots, X_n as in the previous problem set $T_j = X_j/(X_0 + X_1 + \dots + X_n)$ for $0 \leq j \leq n - 1$. Show that T_0, T_1, \dots, T_{n-1} have the joint Dirichlet density

$$\frac{\Gamma(m_0 + m_1 + \dots + m_n)}{\Gamma(m_0)\Gamma(m_1)\dots\Gamma(m_n)} t_0^{m_0-1} t_1^{m_1-1} \dots t_{n-1}^{m_{n-1}-1} (1 - t_0 - t_1 - \dots - t_{n-1})^{m_n-1}.$$

[Hint: This looks formidable but by setting $S = X_0 + X_1 + \dots + X_n$, the inverse transformation $x_0 = st_0, x_1 = st_1, \dots, x_{n-1} = st_{n-1}, x_n = s - st_0 - st_1 - \dots - st_{n-1}$ has a simple Jacobian. It is easy hence to write down the (joint) density of $T_0, T_1, \dots, T_{n-1}, S$ whence the result follows by integration over s .]

26. *The Cauchy density in the plane.* Suppose (X_1, X_2) has density $(2\pi)^{-1}(1+x_1^2+x_2^2)^{-3/2}$. By switching to polar coordinates, $X_1 = R \cos \Theta$ and $X_2 = R \sin \Theta$, show that R and Θ are independent.

27. *The Cauchy density in \mathbb{R}^3 .* Show that $f(x_1, x_2, x_3) = [\pi(1+x_1^2+x_2^2+x_3^2)]^{-2}$ is a density and compute its marginals.

The following problems deal with order statistics. Consider the rearrangement of an ordered n -tuple of numbers (x_1, \dots, x_n) in order of increasing size, $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. The numbers $x_{(k)}$ are called the order statistics of (x_1, \dots, x_n) with $x_{(k)}$ being the k th-order statistic. Now suppose X_1, \dots, X_n are independent random variables with common density $f(x)$ and common d.f. $F(x)$, and let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be their order statistics. Suppressing the implicit dependence on n , let $F_{(k)}(x)$ be the distribution function of the k th-order statistic $X_{(k)}$ and let $f_{(k)}(x)$ be its density.

28. *Order statistics for the uniform.* Suppose each X_j is distributed uniformly in the unit interval. Show that $F_{(k)}(x) = \sum_{j=k}^n \binom{n}{j} x^j (1-x)^{n-j}$ for $0 < x < 1$. Hence infer that $f_{(k)}(x) = n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k}$ in the unit interval.

◆ 29. *Continuation, limit theorem for extremal order statistics.* Suppose k is any fixed positive integer and t any positive real value. Show that $F_{(k)}\left(\frac{t}{n}\right) \rightarrow G_k(t)$ as $n \rightarrow \infty$ where G_k is the gamma d.f. of (7.1).

◆ 30. *Continuation, limit theorem for central order statistics.* Suppose $k = k_n$ and $x = x_n$ vary with n such that $nx(1-x) \rightarrow \infty$ and $(nx-k)/\sqrt{nx(1-x)} \rightarrow t$ for some fixed real t as $n \rightarrow \infty$. Show that

$$F_{(k)}(x) = P\{X_{(k)} \leq x\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} ds \quad (n \rightarrow \infty).$$

[Hint: The de Moivre–Laplace limit theorem (see Section VI.5 and Problem VI.11) tells us that the tails of the binomial tend to the normal in a suitable range even when k and x are both functions of n .]

◆ 31. *Continuation, central limit theorem for the sample median.* With $n = 2v - 1$ the v th-order statistic $X_{(v)}$ is then called the *sample median*. Show that $X_{(v)}$ has the limiting distribution

$$P\left\{X_{(v)} \leq \frac{1}{2} + \frac{t}{2\sqrt{2v}}\right\} = F_{(v)}\left(\frac{1}{2} + \frac{t}{2\sqrt{2v}}\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} dt \quad (n \rightarrow \infty).$$

32. *Order statistics for the exponential.* Determine the density of the k th-order statistic if each of the X_j has an exponential distribution with mean $1/\alpha$.

33. *Continuation, increments.* Setting $X_{(0)} = 0$ we may define the increments of the order statistics as the random variables $Y_k = X_{(k)} - X_{(k-1)}$ for $1 \leq k \leq n$. When $n = 2$ show that $P\{Y_1 > t_1, Y_2 > t_2\} = e^{-2\alpha t_1} e^{-\alpha t_2}$ for all positive t_1 and t_2 . Generalise and show that $P\{Y_1 > t_1, \dots, Y_n > t_n\} = e^{-n\alpha t_1} e^{-(n-1)\alpha t_2} \dots e^{-2\alpha t_{n-1}} e^{-\alpha t_n}$ for all positive t_1, \dots, t_n .

34. *Continuation, independence.* Conclude that the order statistics increments Y_1, \dots, Y_n are independent random variables and, for each value of k , the increment Y_k has the marginal exponential density $f_k(x) = (n - k + 1)\alpha e^{-(n - k + 1)\alpha x}$ for $x > 0$.

35. *Continuation, convolution.* Show that $f_{(k)}(x) = (f_1 * \dots * f_k)(x)$.

36. *The n -server queue.* Consider the single-server queue of Section 8 with the refinement that the server (node, transmission line, switch) now can provide service to n customers in parallel. Suppose all service is modelled as independent and exponential with mean $1/\alpha$. By using the memoryless property of the exponential distribution show that the time intervals between successive departures are independent random variables with a common exponential density $n\alpha e^{-n\alpha x}$ ($x > 0$). Hence determine the waiting time distribution of an arriving customer who finds $k > n$ customers ahead of him (including the customers in service).

X

The Coda of the Normal

The work of Carl Friedrich Gauss who in 1809 used the normal density to predict the locations of astronomical objects brought awareness of the density to a wide scientific audience; in consequence it is also called the *Gaussian* density. Its ubiquitous appearance (partly explained by the central limit theorem) has led to its being referred to as the *normal* curve, the popularisation of the nomenclature dating to the influential English statistician Karl Pearson and perhaps also Francis Galton. This is the most important distribution of them all.

C 1–4
A 5, 6, 8, 9
F 7, 10–12

1 The normal density

The *standard normal density* $\phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$ was introduced in the context of the normal limit law in Section VI.1. That it is properly normalised, that is to say, a *density*, follows by Lemma VI.1.1, while by Lemma VI.1.2 it is seen to have mean zero and unit variance so that it is standardised with the proper centring and scale. The associated standard normal distribution function is denoted

$$\Phi(x) := \int_{-\infty}^x \phi(t) dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt.$$

The normal d.f. cannot be written in terms of other elementary functions and should be considered to be an elementary function in its own right like the exponential, logarithmic, and trigonometric functions. While the normal d.f. and density have been extensively tabulated in the past, the simplest scientific calculators now routinely provide numerical estimates of the distribution to very high degrees of precision. Figure 1 shows the graphs of the standard normal density and d.f.

A new centring μ with a scaling of the axis by $\sigma > 0$ yields the general type of the normal density

$$\phi_{\mu, \sigma}(x) = \frac{1}{\sigma} \phi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left\{\frac{-(x - \mu)^2}{2\sigma^2}\right\},$$

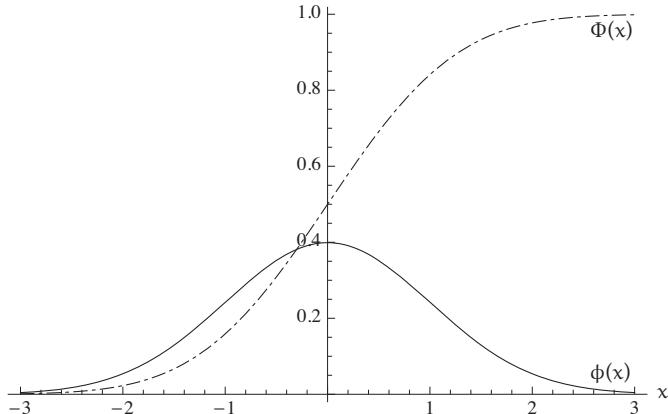


Figure 1: The standard normal density and d.f.

with the corresponding d.f. $\Phi_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$. (In this notation, of course, the standard density and d.f. would become $\phi = \phi_{0,1}$ and $\Phi = \Phi_{0,1}$, respectively. The extended notation is not standard.) It follows quickly that $\phi_{\mu,\sigma}$ has mean μ and variance σ^2 and in consequence *a normal density is completely specified by its mean and variance*. Thus, if X has density $\phi_{\mu,\sigma}$ it is customary to say simply that X is normal with mean μ and variance σ^2 or, even more compactly in notation, to write $X \sim \mathcal{N}(\mu, \sigma^2)$ to mean the same thing. The lucky fact of such a simple parametrisation of the density in terms of just two parameters, the mean and the variance, enormously simplifies calculations involving normals.

Not surprisingly given the prevalence of the normal in theory and applications, a variety of alternative notation has developed for the normal tails in different communities. In statistical communication theory it is usual to express the right tail of the normal as the *Q-function*, $Q(x) = 1 - \Phi(x)$. In the older statistical literature, one also frequently encounters the *error function* defined for $x \geq 0$ by $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$. Again, this is just a simple notational variant as a change of variable shows that $\Phi(x) = \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{x}{\sqrt{2}}\right)$ for $x \geq 0$.

EXAMPLES: 1) *Grading on a curve*. In large classes it is not unusual for the instructor to graft a normal distribution onto the actual distribution of grades. Grade cut-offs are then set at constant multiples of the standard deviation. If the fit to the normal is good and the cut-off for an A is two standard deviations above the mean then the fraction of students getting an A is approximately $\Phi(-2) = .0227 \dots$. Thus, with a 2σ cut-off only about 2% of the students will receive an A. If the cut-off for an A is one standard deviation above the class mean, however, roughly 16% of the students will get an A.

2) *The 3σ rule*. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. The probability that X deviates from its

mean by more than three standard deviations may be estimated by

$$\begin{aligned} \mathbb{P}\{|X - \mu| > 3\sigma\} &= \int_{-\infty}^{\mu-3\sigma} \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dt + \int_{\mu+3\sigma}^{\infty} \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dt \\ &= 2 \int_3^{\infty} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du = 2[1 - \Phi(3)] = 0.0028\cdots. \end{aligned}$$

It follows that to a very good practical approximation the normal may be considered to be concentrated within three standard deviations of its mean. This is the origin of the 3σ rule of thumb that has achieved the status of a folk theorem in applications. ►

Very good estimates are available for the normal tails in view of the rapid quenching of the exponential. For any $x > 0$, the right tail of the standard normal density is given by

$$1 - \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2} dt = \frac{1}{2\sqrt{\pi}} \int_{x^2/2}^{\infty} u^{-1/2} e^{-u} du,$$

the final form obtained via a change of the variable of integration to $u = t^2/2$. The resulting expression is particularly amenable to a successive integration by parts. To begin,

$$\begin{aligned} 1 - \Phi(x) &= \frac{-u^{-1/2} e^{-u}}{2\sqrt{\pi}} \Big|_{x^2/2}^{\infty} - \frac{1}{4\sqrt{\pi}} \int_{x^2/2}^{\infty} u^{-3/2} e^{-u} du \\ &= \frac{e^{-x^2/2}}{x\sqrt{2\pi}} - \frac{1}{4\sqrt{\pi}} \int_{x^2/2}^{\infty} u^{-3/2} e^{-u} du. \end{aligned}$$

The integral on the right is strictly positive so that $1 - \Phi(x) < \frac{\Phi(x)}{x}$. Proceeding, another integration by parts yields

$$\begin{aligned} 1 - \Phi(x) &= \frac{e^{-x^2/2}}{x\sqrt{2\pi}} + \frac{u^{-3/2} e^{-u}}{4\sqrt{\pi}} \Big|_{x^2/2}^{\infty} + \frac{3}{8\sqrt{\pi}} \int_{x^2/2}^{\infty} u^{-5/2} e^{-u} du \\ &= \frac{e^{-x^2/2}}{x\sqrt{2\pi}} - \frac{e^{-x^2/2}}{x^3\sqrt{2\pi}} + \frac{3}{8\sqrt{\pi}} \int_{x^2/2}^{\infty} u^{-5/2} e^{-u} du. \end{aligned}$$

Again, the integral on the right is strictly positive so that we obtain the lower bound $1 - \Phi(x) > \frac{\Phi(x)}{x} - \frac{\Phi(x)}{x^3}$. As the ratio of the upper to the lower bound tends to one as $x \rightarrow \infty$ we obtain a very accurate asymptotic estimate.

THEOREM 1 (NORMAL TAIL BOUND) *The two-sided inequality for the normal tail*

$$\frac{\Phi(x)}{x} \left(1 - \frac{1}{x^2}\right) < 1 - \Phi(x) < \frac{\Phi(x)}{x}$$

holds for each $x > 0$. In particular, $1 - \Phi(x) \sim \frac{1}{x} \Phi(x)$ asymptotically as $x \rightarrow \infty$.

As the normal density is symmetric about the mean, $1 - \Phi(x) = \Phi(-x)$, we may write equivalently $\Phi(-x) \sim \frac{1}{x}\phi(x)$ as $x \rightarrow \infty$. The asymptotic estimate is remarkably accurate even for small values of x . An alternative, essentially geometric, bounding technique utilising the circular symmetry implicit in the normal appears in Lemma VI.1.3.

Suppose $X_1, X_2, \dots, X_n, \dots$ are independent, normal random variables. For each j , suppose X_j has mean μ_j and variance σ_j^2 . For each positive integer n , let $S_n = X_1 + \dots + X_n$. The density of S_n is then given by the n -fold convolution $\phi_{\mu_1, \sigma_1} * \dots * \phi_{\mu_n, \sigma_n}$.

We begin, as usual, by a consideration of the case $n = 2$. We may centre the variables and, by a common scaling, suppose that one of them has unit variance. Accordingly, set $Z_1 = (X_1 - \mu_1)/\sigma_1$ and $Z_2 = (X_2 - \mu_2)/\sigma_1$. Then, Z_1 and Z_2 are independent and normally distributed, Z_1 has mean zero and variance one, and Z_2 has mean zero and variance $\sigma_2^2 = \sigma_2^2/\sigma_1^2$. The sum $Z_1 + Z_2$ hence has density given by the convolution $(\phi_{0,1} * \phi_{0,\sigma})(t)$ which, by collecting exponents and setting $q_t(x) = (t - x)^2 + x^2/\sigma^2$, we may write in the form $\int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp(-\frac{1}{2}q_t(x)) dx$. Temporarily write $m = t\sigma^2/(1 + \sigma^2)$, $s^2 = \sigma^2/(1 + \sigma^2)$, and $a^2 = 1 + \sigma^2$. Completing the square in the exponent and rearranging terms we obtain $q_t(x) = (x - m)^2/s^2 + t^2/a^2$ and as $2\pi\sigma = (\sqrt{2\pi}a)(\sqrt{2\pi}s)$, we may factor the integrand $\frac{1}{2\pi\sigma} \exp(-\frac{1}{2}q_t(x))$ into the product $\phi_{0,a}(t)\phi_{m,s}(x)$. Matters now simplify dramatically leading to

$$(\phi_{0,1} * \phi_{0,\sigma})(t) = \phi_{0,a}(t) \int_{-\infty}^{\infty} \phi_{m,s}(x) dx = \phi_{0,a}(t).$$

The nuisance parameters m and s have disappeared from the equation and the distribution is determined solely by the parameter a . Thus, $Z_1 + Z_2$ is normal with mean zero and variance a^2 . As $S_2 = \sigma_1(Z_1 + Z_2) + (\mu_1 + \mu_2)$ is obtained by a simple shift and scale from $Z_1 + Z_2$, it follows that S_2 is normal with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 a^2 = \sigma_1^2 + \sigma_2^2$.

The case of a general sum now follows quickly. We claim that S_n is normal with mean $\mu_1 + \dots + \mu_n$ and variance $\sigma_1^2 + \dots + \sigma_n^2$. Indeed, suppose as induction hypothesis that this is true for some n . As S_n and X_{n+1} are independent it then follows that $S_{n+1} = S_n + X_{n+1}$ is normal with mean $(\mu_1 + \dots + \mu_n) + \mu_{n+1}$ and variance $(\sigma_1^2 + \dots + \sigma_n^2) + \sigma_{n+1}^2$, completing the induction. This is the important *stability theorem*. (For an indirect proof, see Problem VI.4.)

THEOREM 2 (STABILITY OF THE NORMAL DENSITY) Suppose X_1, \dots, X_n are independent and, for each k , X_k is normal with mean μ_k and variance σ_k^2 . Then $S_n = X_1 + \dots + X_n$ is normal with mean $\mu_1 + \dots + \mu_n$ and variance $\sigma_1^2 + \dots + \sigma_n^2$. Or, in brief, normal densities are stable under convolution.

Thus, for the family of normal densities the type is stable under convolution just as for the Cauchy family. Type stability under convolution, as

we shall see, is a particular case of a much more general stability property that uniquely characterises the normal.

2 Squared normals, the chi-squared density

A variety of distributions important in statistics arise out of the normal. Many of these deal with the *squares* of normal random variables.

Suppose X_1, X_2, \dots, X_n are independent, standard normal random variables. Write $V^2 = X_1^2 + \dots + X_n^2$. What can be said about its density?

For $n = 1$ the situation is very simple. In this case $V^2 = X_1^2$ is immediately seen to have density $\frac{1}{2\sqrt{t}} \{\phi(\sqrt{t}) + \phi(-\sqrt{t})\} = \frac{1}{\sqrt{2\pi t}} e^{-t/2}$ for $t > 0$. In the notation of (IX.8.1), after a little massaging the right-hand side stands revealed as just the gamma density $g_{1/2}(t; 1/2)$. It follows that the random variables X_1^2, \dots, X_n^2 are independent and share the common gamma density $g_{1/2}(\cdot; 1/2)$.

It follows immediately by virtue of the closure of the gamma densities under convolution (see Section IX.8) that, for each positive integer n , the variable $V^2 = X_1^2 + \dots + X_n^2$ has the gamma density

$$(g_{1/2}(\cdot; 1/2) * \dots * g_{1/2}(\cdot; 1/2))(t) = g_{n/2}(t; 1/2),$$

a fact worth recording because of its frequent recurrence in applications.

THEOREM Suppose X_1, \dots, X_n are independent, standard normal random variables. Then the positive variable $V^2 = X_1^2 + \dots + X_n^2$ has density

$$g_{n/2}(t; 1/2) = \frac{1}{2^{n/2}\Gamma(n/2)} t^{n/2-1} e^{-t/2} \quad (t > 0).$$

For historical reasons dating to Karl Pearson this is called the *chi-squared density with n degrees of freedom*. For applications in statistics see Sections XX.10–12.

EXAMPLES: 1) *Folded normal.* For $n = 1$, the positive square-root $U = \sqrt{V^2}$ has density $2tg_{1/2}(t^2; 1/2) = (\frac{2}{\pi})^{1/2} e^{-t^2/2}$ which corresponds to simply reflecting the portion of the normal density on the negative half-axis onto the positive half-axis. As $\sqrt{V^2} = |X_1|$ this should occasion no great surprise.

2) *Return to exponentials.* For $n = 2$, the sum $V^2 = X_1^2 + X_2^2$ has density $g_1(t; \frac{1}{2}) = \frac{1}{2} e^{-t/2}$ concentrated on the positive half-axis. This is just the ordinary exponential density with mean 2. The normalised variable $\frac{1}{2}(X_1^2 + X_2^2)$ represents a scale of V^2 by the factor 1/2 and hence has exponential density with mean 1.

3) *Fading in cellular radio.* Users of cellular phones are familiar with a bane of mobile telephony where a call abruptly fades or is lost. This pernicious effect

can be modelled as due to a random, time-varying, multiplicative term that affects the amplitude of the transmitted wavefront. This effect is caused by the presence of electromagnetic scatterers in the environment that create multiple paths between transmitter and receiver; the phenomenon is particularly pronounced in urban environments where tall buildings provide a multitude of reflecting surfaces. Waveforms from the various scatterers arrive with varying delays and amplitudes at the receiver and can destructively interfere causing an abrupt loss in a call's amplitude. If there are very many independent scatterers the fading process can be modelled (*vide* the central limit theorem) as contributing a normal component, say X_1 , in phase with the transmitted wavefront and an independent normal component, say X_2 , that is out of phase with the transmitted wavefront. These together result in a modulation of the amplitude of the transmitted waveform by the multiplicative term $R = (X_1^2 + X_2^2)^{1/2}$. Of course, as the vehicle moves the scattering landscape changes constantly creating a time-varying random modulation of the amplitude. As I said, pernicious.

If we may model X_1 and X_2 at any given instant as independent, standard normal random variables then R has density $2t g_1(t^2; 1/2) = te^{-t^2/2}$ concentrated on the positive half-axis $t > 0$. This is *Rayleigh's density*.

Fades of this type model situations where there is no direct line of sight between transmitter and receiver. When a direct line of sight is available, the variables X_1 and X_2 are modelled by independent normal random variables of *non-zero* means, say μ_1 and μ_2 , and common variance σ^2 . This leads to *Rice's density* for the fade amplitude $R = (X_1^2 + X_2^2)^{1/2}$; see Problem 4.

4) *Maxwell's distribution of velocities.* Suppose X_1 , X_2 , and X_3 denote the projections along the three coordinate axes of the velocity of a particle in ordinary three-dimensional space. If the X_i are modelled as independent, standard normal random variables then $V^2 = X_1^2 + X_2^2 + X_3^2$ is the square of the speed of the particle with density $g_{3/2}(t; 1/2) = (2\pi)^{-1/2} t^{1/2} e^{-t/2}$. The speed of the particle, $V = (X_1^2 + X_2^2 + X_3^2)^{1/2}$ hence has density $2t g_{3/2}(t^2; 1/2) = (\frac{2}{\pi})^{1/2} t^2 e^{-t^2/2}$ for $t > 0$. This is *Maxwell's density*. ▶

3 A little linear algebra

Vector and matrix notation helps simplify the passage to more than one dimension. The reader familiar with linear algebra should, after a brief glance at our notational conventions, pass on to the next section.

As before, we use bold font notation to denote vectors, $\mathbf{x} = (x_1, \dots, x_n)$ denoting a generic point in \mathbb{R}^n . In keeping with the conventions introduced in Section VII.5, we interpret \mathbf{x} as a *row vector* in matrix–vector operations with its transpose \mathbf{x}^\top representing the associated column vector. As usual, the *inner product* of two vectors \mathbf{x} and \mathbf{y} is given by $\mathbf{x}\mathbf{y}^\top = x_1y_1 + \dots + x_ny_n$ and the

norm (or length) of any vector \mathbf{x} is given by $\|\mathbf{x}\| = \sqrt{\mathbf{x}\mathbf{x}^\top} = \sqrt{x_1^2 + \cdots + x_n^2}$. We say that two vectors \mathbf{x} and \mathbf{y} are *orthogonal* if $\mathbf{x}\mathbf{y}^\top = 0$.

A real square matrix $\mathbf{A} = [a_{jk}]$ of order n induces a linear transformation on \mathbb{R}^n via the map $\mathbf{x} \mapsto \mathbf{x}\mathbf{A}$ which results in a vector $\mathbf{y} = \mathbf{x}\mathbf{A}$ with components $y_k = \sum_{k=1}^n x_j a_{jk}$. Of particular interest are the subspaces of \mathbb{R}^n that \mathbf{A} maps back into themselves.¹ We say that a subspace \mathbb{S} of \mathbb{R}^n is *invariant under \mathbf{A}* if $\mathbf{x}\mathbf{A} \in \mathbb{S}$ for every $\mathbf{x} \in \mathbb{S}$. Of course, \mathbb{R}^n itself is an (n -dimensional) invariant subspace. An important fact is that there exist non-trivial invariant subspaces of smaller dimensionality.

THEOREM 1 *Every linear transformation on \mathbb{R}^n has a one-dimensional or a two-dimensional invariant subspace.*

PROOF: Adopt the nonce notation \mathbf{I} for the identity matrix of order n . The vector equation $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$ or, equivalently, the homogeneous system of equations

$$\begin{aligned} x_1 a_{11} + x_2 a_{21} + \cdots + x_n a_{n1} &= \lambda x_1, \\ x_1 a_{12} + x_2 a_{22} + \cdots + x_n a_{n2} &= \lambda x_2, \\ \dots & \\ x_1 a_{1n} + x_2 a_{2n} + \cdots + x_n a_{nn} &= \lambda x_n, \end{aligned} \tag{3.1}$$

has a non-zero solution if, and only if, the matrix $\mathbf{A} - \lambda\mathbf{I}$ is singular. Writing

$$p(\lambda) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix}, \tag{3.2}$$

it is clear by expanding the determinant on the right that $p(\lambda)$ is a polynomial of degree n in the real variable λ —this is called the *characteristic polynomial of \mathbf{A}* . The solutions of (3.1) are hence determined by the roots of the characteristic polynomial. Now, by the fundamental theorem of algebra, any non-constant polynomial has a real or complex root. We consider the cases in turn.

Suppose λ is a real root of the characteristic polynomial (3.2). Then there exists a real, non-zero solution $\mathbf{x} = (x_1, \dots, x_n)$ for the vector equation $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$. The vector \mathbf{x} spans a one-dimensional subspace of rays $c\mathbf{x}$ obtained by scaling \mathbf{x} and as it is clear that $(c\mathbf{x})\mathbf{A} = \lambda(c\mathbf{x})$ for every real c , it follows that the family of rays $\{c\mathbf{x} : c \in \mathbb{R}\}$ is a one-dimensional invariant subspace of \mathbf{A} .

Suppose alternatively that $\lambda = \alpha + i\beta$ is a complex root of the characteristic polynomial, $\beta \neq 0$, and let $x_1 = \xi_1 + i\eta_1, \dots, x_n = \xi_n + i\eta_n$ be

¹A subspace \mathbb{S} of \mathbb{R}^n is a subset of \mathbb{R}^n that is closed under vector addition and scalar multiplication: if \mathbf{x} and \mathbf{y} are elements of \mathbb{S} and λ is any real constant then $\mathbf{x} + \mathbf{y}$ and $\lambda\mathbf{x}$ are also elements of \mathbb{S} .

a corresponding solution of (3.1). By separating real and imaginary parts, we obtain the twin system of equations

$$\begin{array}{ll} \xi_1 a_{11} + \cdots + \xi_n a_{n1} = \alpha \xi_1 - \beta \eta_1 & \eta_1 a_{11} + \cdots + \eta_n a_{n1} = \beta \xi_1 + \alpha \eta_1 \\ \xi_1 a_{12} + \cdots + \xi_n a_{n2} = \alpha \xi_2 - \beta \eta_2 & \eta_1 a_{12} + \cdots + \eta_n a_{n2} = \beta \xi_2 + \alpha \eta_2 \\ \dots & \dots \\ \xi_1 a_{1n} + \cdots + \xi_n a_{nn} = \alpha \xi_n - \beta \eta_n & \eta_1 a_{1n} + \cdots + \eta_n a_{nn} = \beta \xi_n + \alpha \eta_n \end{array}$$

which, by setting $\mathbf{y} = (\xi_1, \dots, \xi_n)$ and $\mathbf{z} = (\eta_1, \dots, \eta_n)$, we may write compactly in the form

$$\mathbf{y}\mathbf{A} = \alpha\mathbf{y} - \beta\mathbf{z} \quad \text{and} \quad \mathbf{z}\mathbf{A} = \beta\mathbf{y} + \alpha\mathbf{z}. \quad (3.3)$$

If $\mathbf{x} = a\mathbf{y} + b\mathbf{z}$ is in the linear span of \mathbf{y} and \mathbf{z} then so is $\mathbf{x}\mathbf{A} = a\mathbf{y}\mathbf{A} + b\mathbf{z}\mathbf{A} = (a\alpha + b\beta)\mathbf{y} + (b\alpha - a\beta)\mathbf{z}$. It follows that the set $\{a\mathbf{y} + b\mathbf{z} : a, b \in \mathbb{R}\}$ is invariant under \mathbf{A} ; this is the two-dimensional subspace spanned by \mathbf{y} and \mathbf{z} . ▶

One-dimensional invariant subspaces are of particular importance. To each one-dimensional invariant subspace of \mathbf{A} is associated a real value λ such that $\mathbf{x}\mathbf{A} = \lambda\mathbf{x}$ for any \mathbf{x} in the (one-dimensional) subspace. Any element \mathbf{x} of such a subspace is called an *eigenvector* of \mathbf{A} , λ the corresponding *eigenvalue*.

The irritating possibility that \mathbf{A} has no one-dimensional invariant subspace is avoided by providing \mathbf{A} with some more structure. Of particular interest to us is the case when \mathbf{A} is *symmetric* (also called *self-adjoint*) in which case we may identify \mathbf{A} with its transpose, $\mathbf{A} = \mathbf{A}^T$ (or, equivalently, $a_{jk} = a_{kj}$ for all j and k). In this case \mathbf{A} is equipped with a full complement of one-dimensional invariant subspaces and more besides. We begin by exhibiting that *one* one-dimensional invariant subspace exists.

LEMMA 1 *Every symmetric linear transformation on a finite-dimensional space has a one-dimensional invariant subspace.*

PROOF: Suppose \mathbf{A} is a symmetric real matrix of order n and suppose $\lambda = \alpha + i\beta$ is a complex root of the characteristic polynomial. Let \mathbf{y} and \mathbf{z} be as in (3.3). As $\mathbf{A} = \mathbf{A}^T$, we have $(\mathbf{y}\mathbf{A})\mathbf{z}^T = \mathbf{y}\mathbf{A}\mathbf{z}^T = \mathbf{y}\mathbf{A}^T\mathbf{z}^T = \mathbf{y}(\mathbf{z}\mathbf{A})^T$. On the other hand, by two applications of (3.3), we have $(\mathbf{y}\mathbf{A})\mathbf{z}^T = \alpha\mathbf{y}\mathbf{z}^T - \beta\mathbf{z}\mathbf{z}^T$ and $\mathbf{y}(\mathbf{z}\mathbf{A})^T = \beta\mathbf{y}\mathbf{y}^T + \alpha\mathbf{y}\mathbf{z}^T$. Equating the two expressions we obtain $\beta(\mathbf{y}\mathbf{y}^T + \mathbf{z}\mathbf{z}^T) = 0$ and as the term inside the parentheses is the sum of the squares of the lengths of the vectors \mathbf{y} and \mathbf{z} , hence strictly positive, it follows that $\beta = 0$ identically. The characteristic polynomial of \mathbf{A} hence has a real root λ which implies in turn the existence of a one-dimensional invariant subspace. ▶

LEMMA 2 *Suppose \mathbf{A} is a symmetric linear transformation on \mathbb{R}^n , \mathbf{u}_1 an eigenvector of \mathbf{A} with associated eigenvalue λ_1 . Then the $(n - 1)$ -dimensional subspace of vectors orthogonal to \mathbf{u}_1 is invariant under \mathbf{A} .*

PROOF: Suppose that v is orthogonal to u_1 , $vu_1^\top = 0$. As A is symmetric, we have $(vA)u_1^\top = vAu_1^\top = vA^\top u_1^\top = v(u_1 A)^\top = \lambda_1 vu_1^\top = 0$. \blacktriangleright

The stage is set for an induction. Suppose as in Lemma 2 that $u_1 = (u_{11}, \dots, u_{1n})$ is an eigenvector of the symmetric matrix A , λ_1 the corresponding eigenvalue. We now attempt to solve the system (3.1) subject to the constraint that x lies in the $(n - 1)$ -dimensional subspace orthogonal to u_1 , that is to say, $xu_1^\top = x_1u_{11} + \dots + x_nu_{1n} = 0$. As $u_1 \neq 0$, it has at least one non-zero component. Suppose, without loss of generality, that $u_{1n} \neq 0$. Then $x_n = -\frac{1}{u_{1n}}(x_1u_{11} + \dots + x_{n-1}u_{1,n-1})$ and by eliminating x_n from the system (3.1) we obtain a system of $n - 1$ equations in $n - 1$ unknowns with symmetric weighting coefficients. The corresponding characteristic polynomial has degree $n - 1$ and, again, by the fundamental theorem of algebra, has a real or complex root. As the weights are symmetric, by Lemma 1 it follows again that the root must be real. There hence exists an eigenvector u_2 , with corresponding real eigenvalue λ_2 , in the subspace orthogonal to u_1 . Proceeding inductively in this fashion we may generate a system of n mutually orthogonal eigenvectors u_1, \dots, u_n of A . We may select for these unit-length vectors along each of the one-dimensional invariant subspaces. These vectors are orthogonal, *a fortiori* linearly independent, and hence form the basis of an orthogonal coordinate system for \mathbb{R}^n . This is a fact of great importance.

THEOREM 2 *Every symmetric linear transformation on \mathbb{R}^n engenders an orthonormal basis of eigenvectors.*

We may tease a little more information out of the framework. If u_1 and u_2 are eigenvectors of A with distinct eigenvalues $\lambda_1 \neq \lambda_2$, then $\lambda_1 u_1 u_2^\top = (u_1 A)u_2^\top = u_1(u_2 A)^\top = \lambda_2 u_1 u_2^\top$, whence $u_1 u_2^\top = 0$. Thus, *eigenvectors corresponding to distinct eigenvalues are orthogonal*. In the case of repeated eigenvalues suppose $\lambda = \lambda_{j_1} = \dots = \lambda_{j_k}$. Then any vector $x = c_1 u_{j_1} + \dots + c_k u_{j_k}$ in the linear span of the associated eigenvectors is also an eigenvector of A with eigenvalue λ as is easy to see: $xA = c_1 \lambda_{j_1} u_{j_1} + \dots + c_k \lambda_{j_k} u_{j_k} = \lambda x$. It follows that *eigenvectors corresponding to a common eigenvalue span an invariant subspace of A* .

Matrix notation simplifies equations and brings the salient fact emphasised by Theorem 2 into sharp relief. Write

$$U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \dots & \dots & \dots & \dots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{pmatrix} \quad (3.4)$$

for the $n \times n$ matrix whose rows are orthonormal eigenvectors of A and let $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ be the diagonal matrix of corresponding eigenvalues. We may then write the n eigenvalue equations $u_1 A = \lambda_1 u_1$, $u_2 A = \lambda_2 u_2$, ... ,

$\mathbf{u}_n \Lambda = \lambda_n \mathbf{u}_n$ compactly in the form $U\Lambda = \Lambda U$. By construction, the matrix U is orthogonal, $U^T U = I$ or, equivalently, $U^T = U^{-1}$. By pre-multiplying or post-multiplying both sides of the matrix eigenvalue equation by U^T we hence obtain the equivalent representations $\Lambda = U^T \Lambda U$ and $U \Lambda U^T = \Lambda$, or, in words, *the matrix Λ is diagonalisable with respect to an orthonormal basis of eigenvectors.*

As U is orthogonal, $|\det(U)| = |\det(U^T)| = 1$ (if the reader does not know this fact she can verify it by taking determinants of both sides of the identity $U^T U = I$). By taking determinants in the matrix eigenvalue equation we hence obtain $\det(\Lambda) = \det(\Lambda) = \lambda_1 \lambda_2 \cdots \lambda_n$. Thus, Λ is non-singular if, and only if, none of its eigenvalues is zero. An important special case arises when all the eigenvalues of Λ are strictly positive in which case we say that Λ is *strictly positive definite*. By inverting the eigenvalue equation it is then clear that Λ^{-1} is also strictly positive definite with the same eigenvectors as Λ , the corresponding eigenvalues being $\lambda_1^{-1}, \dots, \lambda_n^{-1}$. It follows that $\Lambda^{-1} = U^T \Lambda^{-1} U$ and we obtain a diagonalisation of Λ^{-1} with respect to the same orthonormal basis.

4 The multivariate normal

Our experience with the univariate and bivariate normals suggests that a normal density in n dimensions should be characterised, if at all, by a quadratic form in the exponent. Suppose accordingly that $Q = [q_{jk}]$ is an $n \times n$ symmetric matrix, $q_{jk} = q_{kj}$. Associated with Q is the quadratic form q which maps each $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ into the real value $q(\mathbf{x}) = \mathbf{x} Q \mathbf{x}^T = \sum_{j,k=1}^n q_{jk} x_j x_k$.

DEFINITION We say that a density f in n dimensions is a *normal density centred at the origin* if it is of the form $f(\mathbf{x}) = C e^{-q(\mathbf{x})/2}$. The shift of origin $f(\mathbf{x} - \mathbf{m})$ yields a *normal density centred at the point $\mathbf{m} = (m_1, \dots, m_n)$* .

The factor $1/2$ in the exponent is dictated purely for reasons of algebraic convenience as the term $q_{jk} x_j x_k = q_{kj} x_k x_j$ is repeated in the quadratic form. As f is a density it must be properly normalised and so the integral $\int_{\mathbb{R}^n} e^{-q(\mathbf{x})/2} d\mathbf{x}$ converges; the quantity $C = C_n$ is the corresponding normalisation constant.

The existence of normal densities in dimensions > 2 is seen by a consideration of independent variables. Suppose X_1, \dots, X_n are independent random variables with each $X_j \sim \mathcal{N}(0, \sigma_j^2)$ normal with mean zero and variance σ_j^2 . Then $\mathbf{X} = (X_1, \dots, X_n)$ has density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \sigma_1 \cdots \sigma_n} \exp\left\{-\frac{1}{2}\left(\frac{x_1^2}{\sigma_1^2} + \cdots + \frac{x_n^2}{\sigma_n^2}\right)\right\}.$$

This is the normal density centred at the origin conferred by the diagonal matrix $Q = \text{diag}(\sigma_1^{-2}, \dots, \sigma_n^{-2})$ with strictly positive diagonal entries.

It is too much to hope that arbitrary quadratic forms engender normal densities and we hence turn our attention to a consideration of the conditions

on Q for which a normal density obtains. Our experience with the bivariate normal in Example VII.5.5 suggests that completing squares may be profitable.

Focusing, for definiteness, on the n th coordinate, we may rearrange terms in the quadratic form and write

$$q(\mathbf{x}) = q_{nn}x_n^2 + 2x_n \sum_{j=1}^{n-1} q_{jn}x_j + \sum_{j,k=1}^{n-1} q_{jk}x_jx_k.$$

It is now apparent that q_{nn} cannot be zero as then, with x_1, \dots, x_{n-1} fixed, the expression on the right is of the form $ax_n + b$ and the integral of $e^{-(ax_n+b)/2}$ diverges. Accordingly, we may suppose $q_{nn} \neq 0$. But then we may, by completing the square in x_n , express the quadratic form as $q(\mathbf{x}) = q_{nn}(x_n + \tilde{l})^2 + \tilde{q}$ where \tilde{l} is a linear form and \tilde{q} a quadratic form, both functions of x_1, \dots, x_{n-1} only. With x_1, \dots, x_{n-1} fixed, the integral of $\exp\left\{-\frac{q_{nn}}{2}(x_n + \tilde{l})^2\right\}$ diverges if $q_{nn} < 0$ and converges if $q_{nn} > 0$. As the choice of coordinate direction n is not special, we come to our first conclusion: *the diagonal elements of Q must be strictly positive, $q_{jj} > 0$ for $j = 1, \dots, n$.*

Continuing to focus on the n th coordinate, we may now suppose that $q_{nn} > 0$. Fixing x_1, \dots, x_{n-1} , as before, we write $q(x_1, \dots, x_{n-1}, x_n) = q_{nn}(x_n + \tilde{l}(x_1, \dots, x_{n-1}))^2 + \tilde{q}(x_1, \dots, x_{n-1})$. The expression $e^{-q_{nn}(x_n + \tilde{l})^2/2}$ represents a normal density on the real line with mean $-\tilde{l}$ and variance q_{nn}^{-1} and, consequently,

$$\int_{\mathbb{R}} \exp\left(-\frac{q_{nn}}{2}(x_n + \tilde{l})^2\right) dx_n = \sqrt{\frac{2\pi}{q_{nn}}},$$

the expression on the right invariant with respect to $\tilde{l} = \tilde{l}(x_1, \dots, x_{n-1})$. It follows that by integrating out the variable x_n in the density $f(x_1, \dots, x_{n-1}, x_n) = Ce^{-q(x_1, \dots, x_{n-1}, x_n)/2}$ we are left with an expression in $n - 1$ variables of the form $\tilde{f}(x_1, \dots, x_{n-1}) = \tilde{C}e^{-\tilde{q}(x_1, \dots, x_{n-1})/2}$ where the exponent \tilde{q} is a quadratic form in x_1, \dots, x_{n-1} and \tilde{C} is a strictly positive constant. This function must integrate to 1 over \mathbb{R}^{n-1} (as f is a density) and it follows that \tilde{f} represents a normal density in $n - 1$ dimensions. Another integration results in yet another normal density and, proceeding inductively in this fashion, we come to our next conclusion worthy of enshrining formally.

THEOREM 1 *The marginals of a multivariate normal density are all normal.*

The reader should beware the tempting converse which is false in general: *a system of variables with marginal normal densities need not have a (jointly) normal density, and, in fact, a joint density need not even exist.* See Problem 2. Thus, when we say that a system of variables is normal we always mean that they are *jointly* normal (in the sense that they are described by a density with a quadratic

form in the exponent) and make explicit particular provision if they are merely marginally normal.

We get more insight into the nature of the multivariate normal by a consideration of linear transformations. Suppose $\mathbf{X} = (X_1, \dots, X_n)$ is normal with density $f(\mathbf{x}) = Ce^{-\mathbf{x}^T Q \mathbf{x} / 2} = Ce^{-q(\mathbf{x})/2}$. Suppose W is a non-singular square matrix of order n . A linear transformation of the coordinate variables $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{X}W$ now results in a new system of random variables $\mathbf{Y} = (Y_1, \dots, Y_n)$ with density (see Section VII.9)

$$g(\mathbf{y}) = |\det(W^{-1})| \cdot f(\mathbf{y}W^{-1}) = C|\det W|^{-1} \exp\left\{-\frac{1}{2}\mathbf{y}^T W^{-1} Q (W^{-1})^T \mathbf{y}\right\}.$$

As $W^{-1}Q(W^{-1})^T$ represents another symmetric matrix, we discover another quadratic form in the exponent and g must perform be a normal density.

THEOREM 2 *Non-degenerate linear transformations of normal coordinate variables result in new normal variables.*

Again, the reader should bear in mind that normality in the context of systems of variables refers to joint normality. The theorem is false if the coordinate variables are merely marginally normal. The stability of the normal under convolution may now be viewed through the prism of a much more general stability property: *normals are stable under linear transformations*. In particular, Theorems 1 and 2 together imply that if X_1, \dots, X_n is a (jointly) normal system and $Y = a_1X_1 + \dots + a_nX_n$ then Y is normal and hence has distribution completely determined by its mean and variance. What is key about this observation is that the mean and variance of Y can be determined directly by simple additivity properties.

EXAMPLE: Suppose (X_1, X_2) is a random pair and $Y = X_1 + X_2$. If (X_1, X_2) is governed by the bivariate normal density $\phi(x_1, x_2; \rho)$ of Example VII.5.5 then, as seen, X_1 and X_2 share a common marginal normal distribution with zero mean and unit variance and, moreover, $\text{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) = \rho$. By additivity of expectation, $\mathbb{E}(Y) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 0$, while,

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}((X_1 + X_2)^2) = \mathbb{E}(X_1^2 + 2X_1X_2 + X_2^2) \\ &= \mathbb{E}(X_1^2) + \mathbb{E}(X_2^2) + 2\mathbb{E}(X_1X_2) = 2 + 2\rho,\end{aligned}$$

by another application of additivity. In view of Theorems 1 and 2 it follows without further calculation that $Y \sim \mathcal{N}(0, 2(1 + \rho))$. ▶

What else can the bivariate normal teach us? Example VII.9.2 suggests the beguiling possibility that independent normal variables may be uncovered by a suitable linear transformation of a dependent multivariate normal system. Indeed, consider the linear transformation $\mathbf{x} \mapsto \mathbf{y} = \mathbf{x}W$ defined by $y_1 = x_1$,

$\dots, y_{n-1} = x_{n-1}, y_n = q_{1n}x_1 + \dots + q_{nn}x_n$. It is clear that this transformation is not singular as $\det(W) = q_{nn} > 0$ and so \mathbf{Y} is a normal vector. By completing squares we now see that $q(x_1, \dots, x_n) = \frac{y_n^2}{q_{nn}} + \tilde{q}(y_1, \dots, y_{n-1})$ where \tilde{q} is a quadratic form in y_1, \dots, y_{n-1} . It follows that the density g of \mathbf{Y} may be factored into the product of two normal densities

$$g(y_1, \dots, y_{n-1}, y_n) = \frac{1}{\sqrt{q_{nn}}} \phi\left(\frac{y_n}{\sqrt{q_{nn}}}\right) \cdot \tilde{C} e^{-\tilde{q}(y_1, \dots, y_{n-1})/2}.$$

The random variable Y_n is hence (marginally) normal with zero mean and independent of the normal vector (Y_1, \dots, Y_{n-1}) . Repeating the procedure allows us to peel off independent, zero-mean normal variables, one at a time.

THEOREM 3 *Suppose \mathbf{X} has a normal density centred at the origin in n dimensions. There then exists a non-singular square matrix V of order n such that $\mathbf{Z} = \mathbf{X}V$ is a normal vector with independent coordinates centred at the origin.*

PROOF: As induction hypothesis, we suppose that there exists a non-singular linear transformation $V': (Y_1, \dots, Y_{n-1}) \mapsto (Z_1, \dots, Z_{n-1})$ which transforms the normal vector (Y_1, \dots, Y_{n-1}) into a normal vector (Z_1, \dots, Z_{n-1}) where the variables Z_1, \dots, Z_{n-1} are independent and centred at the origin. As Y_n is independent of (Y_1, \dots, Y_{n-1}) it is also independent of (Z_1, \dots, Z_{n-1}) . Setting $Z_n = Y_n$ and $W' = \begin{bmatrix} V' & 0^\top \\ 0 & 1 \end{bmatrix}$ (where 0 represents the zero vector in $n-1$ dimensions) we find that $\mathbf{Z} = \mathbf{Y}W' = \mathbf{X}WW'$ is a normal vector with independent, zero-mean components. Identifying V with the non-singular matrix product WW' completes the proof. ▶

The transformation V is called a *decorrelating transformation*. The existence of such a transformation opens the door to an explicit characterisation of normal densities. For each j , let σ_j^2 denote the variance of Z_j and introduce the diagonal matrix $\Lambda = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. As Z_1, \dots, Z_n are independent normal variables we may write the density of \mathbf{Z} in the form

$$g(z) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-z_j^2/2\sigma_j^2} = \frac{1}{(2\pi)^{n/2} \det(\Lambda)^{1/2}} \exp(-\frac{1}{2} z \Lambda^{-1} z^\top).$$

On the other hand, the vector $\mathbf{Z} = \mathbf{X}V$ is obtained via a linear transformation of coordinates. The density f of \mathbf{X} may hence be written in terms of the density g of \mathbf{Z} via

$$f(x) = |\det(V)| \cdot g(xV) = \frac{|\det(V)|}{(2\pi)^{n/2} \det(\Lambda)^{1/2}} \exp(-\frac{1}{2} x V \Lambda^{-1} V^\top x^\top). \quad (4.1)$$

While we have made progress, the nature of the linear transformation V which appears on the right is a little mysterious and surely warrants further investigation. It is natural to attempt to relate V to Λ .

The systematic reductions integrating out one variable at a time and leading to Theorem 1 show that, at each stage, we obtain a new normal density centred at the origin in one less coordinate; *a fortiori*, each of the marginal normal densities corresponding to the variables X_1, \dots, X_n has zero mean. Writing A for the covariance matrix of \mathbf{X} , we hence have $A = \text{Cov}(\mathbf{X}) = E(\mathbf{X}^T \mathbf{X})$. By additivity of the expectation integral (see Problem VII.16), it follows that $E(\mathbf{Z}^T \mathbf{Z}) = E(V^T \mathbf{X}^T \mathbf{X} V) = V^T E(\mathbf{X}^T \mathbf{X}) V = V^T A V$. On the other hand, as the normal vector \mathbf{Z} centred at the origin has independent coordinates, the component variables Z_1, \dots, Z_n are mutually uncorrelated and we may hence identify the diagonal matrix Λ with the covariance matrix of \mathbf{Z} . As \mathbf{Z} has zero mean this means that we may write $E(\mathbf{Z}^T \mathbf{Z}) = \text{Cov}(\mathbf{Z}) = \Lambda$. By comparing expressions we hence obtain the simple relation

$$\Lambda = V^T A V \quad (4.2)$$

linking the covariance matrices of \mathbf{Z} and \mathbf{X} through the transformation V . By taking determinants of both sides we see that $0 < \sigma_1^2 \sigma_2^2 \cdots \sigma_n^2 = \det(\Lambda) = \det(V)^2 \det(A)$ and we may draw the useful conclusions that $\det(A) > 0$ and $|\det(V)| = \det(\Lambda)^{1/2} / \det(A)^{1/2}$. As V is non-singular, we may also write (4.2) in the forms $A = (V^T)^{-1} \Lambda V^{-1}$ or $A^{-1} = V \Lambda^{-1} V^T$. Substituting into (4.1), we see that the density of \mathbf{X} has a particularly simple form,

$$Ce^{-\mathbf{x}^T Q \mathbf{x} / 2} = f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(A)^{1/2}} \exp(-\frac{1}{2} \mathbf{x}^T A^{-1} \mathbf{x}).$$

By comparing expressions we see that $C = (2\pi)^{-n/2} \det(A)^{-1/2}$ and $Q = A^{-1}$. The form of the general normal density is now obtained by a shift of origin.

THEOREM 4 *A multivariate normal density f in n dimensions is completely specified by its mean vector \mathbf{m} and its covariance matrix A and takes the form*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(A)^{1/2}} \exp(-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T A^{-1} (\mathbf{x} - \mathbf{m})). \quad (4.3)$$

COROLLARY *A normal vector $\mathbf{X} = (X_1, \dots, X_n)$ has independent components if, and only if, the variables are mutually uncorrelated.*

PROOF: If each pair of variables X_j and X_k is uncorrelated then $A = \text{Cov}(\mathbf{X}) = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ is a diagonal matrix with strictly positive diagonal entries. The exponent on the right in (4.3) then becomes

$$(\mathbf{x} - \mathbf{m})^T A^{-1} (\mathbf{x} - \mathbf{m}) = \frac{1}{\sigma_1^2} (x_1 - m_1)^2 + \cdots + \frac{1}{\sigma_n^2} (x_n - m_n)^2,$$

and we may separate the expression for the joint density into a product of marginal normal densities,

$$f(x_1, \dots, x_n) = \frac{1}{\sqrt{2\pi} \sigma_1} e^{-(x_1 - m_1)^2 / 2\sigma_1^2} \times \cdots \times \frac{1}{\sqrt{2\pi} \sigma_n} e^{-(x_n - m_n)^2 / 2\sigma_n^2}.$$

It follows that the variables X_1, \dots, X_n are independent. By reversing the steps of the argument we can conclude conversely that if the variables X_1, \dots, X_n are marginally normal and independent then the covariance matrix is diagonal, whence the variables are mutually uncorrelated. ▶

The corollary is of very great importance as it asserts the remarkable property that the dependency structure of the entire normal system (X_1, \dots, X_n) is determined by the apparently weak pairwise correlative structure of these variables. This property allows us to unlock the nature of the matrices Q that engender normal densities and the special matrix V of Theorem 3, the recursive nature of the construction of which obscures its form and relationship to Q .

Suppose that W is any *decorrelating transform* for which the transformed coordinate vector $Z = XW$ has mutually uncorrelated elements, that is to say, $\Lambda = \text{Cov}(Z)$ is diagonal. Then Z is normal (by Theorem 2) whence, by the preceding corollary, its elements Z_1, \dots, Z_n are independent and marginally normal. As the matrix V of Theorem 3 leads to mutually uncorrelated variables Z_1, \dots, Z_n , the representation (4.2) suggests that we select for V any diagonalising transformation for the symmetric covariance matrix A . In particular, we may select $V = U^T$ where U is the orthogonal matrix given in (3.4) whose rows u_1, \dots, u_n form an orthonormal basis of eigenvectors of A . We may hence identify the diagonal elements of Λ with the eigenvalues of A and as these represent variances *all eigenvalues of the covariance matrix A must be strictly positive*.

THEOREM 5 *A real symmetric matrix Q determines a normal density if, and only if, it is strictly positive definite. In this case the density has covariance A = Q⁻¹ and is given generically by (4.3) for a density centred at the point m.*

A geometric view helps illuminate the drab algebraic picture. By setting $m = 0$ and centring the density (4.3) at the origin, the level sets of the density f , that is to say, collections of points x for which $f(x) = \text{constant}$, satisfy the quadratic equation $xA^{-1}x^T = \text{constant}$. As A is positive definite, hence also A^{-1} , this represents the equation of an ellipsoid in n dimensions, the principal axes of which are aligned with the eigenvectors of A . The linear transformation V of Theorem 3 may now be identified with a *rotation* of coordinates with the new axis system rotated to align with an orthonormal eigenbasis engendered by A . The reader who prefers a picture to the algebra will find one in Figure 2 which shows the level set $f(x_1, x_2, x_3) = 0.01$ for a trivariate normal system (X_1, X_2, X_3) , centred at the origin, with covariance matrix

$$A = \begin{pmatrix} 1 & 0.4 & 0.5 \\ 0.4 & 1 & 0.3 \\ 0.5 & 0.3 & 1 \end{pmatrix}.$$

As noted earlier, a consequence of Theorems 1 and 2 is that any linear combination $Y_1 = X_1w_{11} + \dots + X_nw_{n1}$ of a normal vector X is normal. If

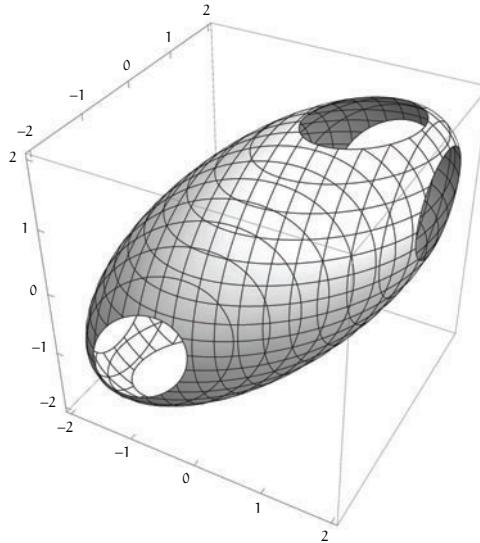


Figure 2: Ellipsoidal level set for a trivariate normal density.

$Y_2 = X_1 w_{12} + \dots + X_n w_{n2}$ then the same is true not only of Y_2 but also of the pair (Y_1, Y_2) provided that Y_1 and Y_2 are not linearly related, that is, they do not satisfy a relationship of the form $c_1 Y_1 + c_2 Y_2 = 0$. If Y_1 and Y_2 have a relationship of this form, however, then the distribution of the pair is concentrated on a line in the plane. In such cases it is desirable to preserve the terminology and say that the pair (Y_1, Y_2) forms a *degenerate normal system* concentrated on a line. More generally, we say that a vector $\mathbf{Y} = (Y_1, \dots, Y_v)$ is *normal* if there exists an n -dimensional normal vector $\mathbf{X} = (X_1, \dots, X_n)$, a point $\mathbf{m} = (m_1, \dots, m_v)$ in v dimensions, and an $n \times v$ real matrix W such that $\mathbf{Y} = \mathbf{m} + \mathbf{X}W$. The distribution of \mathbf{Y} is not degenerate only if $v \leq n$ and the columns of W are linearly independent. A degenerate normal distribution is hence completely specified by a normal distribution on a lower-dimensional manifold of the space and all the considerations of this section carry over without essential change.

5 An application in statistical estimation

Suppose X_1, \dots, X_n are independent random variables drawn from a common normal distribution with mean m and variance σ^2 . In many statistical applications these variables may be thought of as arising by independent sampling from an underlying normal population and, accordingly, X_1, \dots, X_n is called a *random sample*. If the mean m is unknown it is natural to *estimate* it by the arithmetic mean of these variables, $\hat{M} = \frac{1}{n}(X_1 + \dots + X_n)$. In statistics, the variable

\hat{M} is called the *sample mean*. How good is \hat{M} as an estimator of m ?

In view of the stability of the normal under convolution (Theorem 1.2) \hat{M} is normal with mean m (it is an *unbiased* estimator of the mean) and variance σ^2/n . The likely deviations of the sample mean \hat{M} from the true underlying population mean m are hence governed by the distribution of the variable $Z = \frac{\hat{M}-m}{\sigma/\sqrt{n}}$ called the *z-statistic*. As the shift of origin and scale simply normalise Z to zero mean and unit variance, Z is a standard normal variable. Consequently, for every $t > 0$, the tail probabilities $P\{|Z| \leq t\} = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1$ provide *confidence intervals* for the deviation of the estimate \hat{M} from the true (but unknown) m . Thus, in particular,

$$P\{|\hat{M} - m| \leq 2\sigma/\sqrt{n}\} = 2\Phi(2) - 1 = 0.9545$$

and we can assert with a confidence of 95% that the sample mean differs from the population mean by no more than two population standard deviations over the square-root of the sample size. In the language of statistics, the interval $[m - 2\sigma/\sqrt{n}, m + 2\sigma/\sqrt{n}]$ of width $4\sigma/\sqrt{n}$ centred at m is a 95% confidence interval for the estimate of the mean. We have obtained not only an estimate of the unknown m but the size of the likely error.

The reader may well object that while the procedure is all well and good, the specification of, say, a 95% confidence interval for the difference $|\hat{M} - m|$ requires knowledge of the variance σ^2 (as is evident in the construction of the *z*-statistic) and it hardly seems reasonable to assume in general that the variance of the population is known when the mean is not. A principled work-around to this objection is to provide an estimate of the variance as well from the data and it is natural to consider as estimator an expression of the form

$$\hat{\Sigma}^2 = \frac{1}{n-1} [(X_1 - \hat{M})^2 + \dots + (X_n - \hat{M})^2]. \quad (5.1)$$

Statisticians call this the *sample variance*.² We should first of all see what can be said about the distribution of $\hat{\Sigma}^2$.

We may simplify notation by supposing that by a proper centring and scaling, $X_j \leftarrow (X_j - m)/\sigma$, each of the variables X_j has a standard normal density with mean zero and variance one. We then simply have to make the replacements $\hat{M} \leftarrow (\hat{M} - m)/\sigma$ and $\hat{\Sigma}^2 \leftarrow \hat{\Sigma}^2/\sigma^2$ and nothing fundamental has changed in the sample mean and sample variance excepting only a fixed shift and scale.

Consider the vector $Z = (Z_1, \dots, Z_n)$ with components $Z_1 = X_1 - \hat{M}, \dots, Z_n = X_n - \hat{M}$. We then have $(n-1)\hat{\Sigma}^2 = Z_1^2 + \dots + Z_n^2 = \|Z\|^2$. The sample variance is hence identified with the square of the length of Z and inherits its distribution from that of Z . Now, by writing the components $Z_1, \dots,$

²Why divide by $n-1$ and not n ? The reason is that only $n-1$ of the terms $X_1 - \hat{M}, X_2 - \hat{M}, \dots, X_n - \hat{M}$ can be independently specified as $(X_1 - \hat{M}) + (X_2 - \hat{M}) + \dots + (X_n - \hat{M}) = 0$. In the language of statistics, there are only $n-1$ *degrees of freedom*. This concept was introduced by R. A. Fisher. For large n the difference in division by $n-1$ or by n is minute.

Z_n explicitly as a function of the variables X_1, \dots, X_n , it is apparent that Z is obtained from X via the symmetric linear transformation $X \mapsto Z = XW$ where

$$W = \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ \dots & \dots & \dots & \dots & \dots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix}.$$

The transformation is singular as $Z_1 + \dots + Z_n = (X_1 - \hat{M}) + \dots + (X_n - \hat{M}) = 0$. Writing $\mathbf{1} \in \mathbb{R}^n$ for the vector all of whose components are equal to 1, we may write this relation in the form $0 = \mathbf{1}Z^\top = \mathbf{1}W^\top X^\top = \mathbf{1}WX^\top$ as $W = W^\top$. It follows that $\mathbf{1}W = \mathbf{0} = 0 \cdot \mathbf{1}$ (as may also be directly verified from the vector-matrix product) and the vector of all ones is an eigenvector of W with eigenvalue 0. In the language of linear algebra, the one-dimensional invariant subspace corresponding to the ray passing from the origin in \mathbb{R}^n through the point $\mathbf{1} = (1, \dots, 1)$ comprises the *null space* of W . We should now investigate the subspace orthogonal to the null space.

Let \mathbf{q} be any vector in the subspace orthogonal to the null space of W . The j th component of the vector $\mathbf{q}W$ is then given by

$$(\mathbf{q}W)_j = (1 - n^{-1})q_j + \sum_{k \neq j} (-n^{-1})q_k = q_j - n^{-1}q_j - n^{-1} \sum_{k \neq j} q_k = q_j$$

as $\mathbf{q}\mathbf{1}^\top = q_1 + \dots + q_n = 0$. It follows that $\mathbf{q}W = \mathbf{q}$ whence the $(n-1)$ -dimensional subspace orthogonal to the null space of W is an invariant subspace of W corresponding to the eigenvalue 1. Let $\mathbf{q}_n = n^{-1/2}\mathbf{1}$ be a unit vector in the null space of W and select any orthogonal collection of unit vectors $\mathbf{q}_1, \dots, \mathbf{q}_{n-1}$ spanning the $(n-1)$ -dimensional invariant subspace of W orthogonal to \mathbf{q}_n . Then $\mathbf{q}_j W = \mathbf{q}_j$ for $1 \leq j \leq n-1$, $\mathbf{q}_n W = \mathbf{0}$, and $\mathbf{q}_j \mathbf{q}_k^\top$ equals one if $j = k$ and zero otherwise. The vectors $\mathbf{q}_1, \dots, \mathbf{q}_n$ hence form an orthonormal basis of eigenvectors for \mathbb{R}^n . We are now naturally led to consider the transformation $Z \mapsto ZQ$ where Q is the orthogonal matrix whose columns $\mathbf{q}_1^\top, \dots, \mathbf{q}_{n-1}^\top, \mathbf{q}_n^\top$ are the transposes of the selected eigenvectors of W . The transformation Q merely rotates the coordinate system to align with the orthonormal eigenvector basis. Writing $Q' = (\mathbf{q}_1^\top \ \mathbf{q}_2^\top \ \dots \ \mathbf{q}_{n-1}^\top)$ we have $Q = (Q' \ \mathbf{q}_n^\top)$ so that $ZQ = (ZQ' \ n^{-1/2}Z\mathbf{1}^\top) = (Z' \ 0)$ where $Z' = (Z'_1, \dots, Z'_{n-1}) = ZQ'$. Then

$$\|Z\|^2 = ZZ^\top = ZQQ^\top Z^\top = \|ZQ\|^2 = \|(Z' \ 0)\|^2 = \|Z'\|^2 + 0^2 = \|Z'\|^2$$

as Q is orthogonal, whence $QQ^\top = I$; coordinate rotations preserve lengths. And therefore $P\{\|Z\|^2 \leq t\} = P\{\|Z'\|^2 \leq t\}$. By design, the degenerate n th coordinate in the rotated coordinate system plays no further rôle.

The system of variables $\mathbf{Z}' = \mathbf{ZQ}' = \mathbf{XWQ}'$ is obtained by a linear transformation from \mathbf{X} and is consequently normal and centred at the origin. Its density is hence completely determined by its covariance matrix and, once again by additivity of the expectation integral,

$$\text{Cov}(\mathbf{Z}') = \mathbb{E}(\mathbf{Z}'^T \mathbf{Z}') = \mathbb{E}(\mathbf{Q}'^T \mathbf{W}^T \mathbf{X}^T \mathbf{X} \mathbf{W} \mathbf{Q}') = \mathbf{Q}'^T \mathbf{W}^T \mathbb{E}(\mathbf{X}^T \mathbf{X}) \mathbf{W} \mathbf{Q}'.$$

As X_1, \dots, X_n are independent normal variables with zero mean and unit variance, $\text{Cov}(\mathbf{X}) = \mathbb{E}(\mathbf{X}^T \mathbf{X}) = I_n$, the identity matrix of order n . Moreover,

$$\mathbf{WQ}' = \mathbf{W}^T \mathbf{Q}' = (\mathbf{W}^T \mathbf{q}_1^T \quad \mathbf{W}^T \mathbf{q}_2^T \quad \dots \quad \mathbf{W}^T \mathbf{q}_{n-1}^T) = (\mathbf{q}_1^T \quad \mathbf{q}_2^T \quad \dots \quad \mathbf{q}_{n-1}^T) = \mathbf{Q}'.$$

It follows that $\text{Cov}(\mathbf{Z}') = \mathbf{Q}'^T I_n \mathbf{Q}' = \mathbf{Q}'^T \mathbf{Q}' = I_{n-1}$ is the identity matrix in $n-1$ dimensions, that is to say, in the normal system \mathbf{Z}' the component variables Z'_1, \dots, Z'_{n-1} have zero mean, unit variance, and are mutually uncorrelated. By the corollary of the previous section it follows that Z'_1, \dots, Z'_{n-1} are independent, standard normal variables. As $\|\mathbf{Z}'\|^2$ is the sum of the squares of $n-1$ independent normal variables, the theorem of Section 2 shows that it is governed by a chi-squared density. Scaling permits the recovery of the distribution of the sample variance.

THEOREM 1 Let $\hat{\Sigma}^2$ be the sample variance of a random sample of size n drawn by independent sampling from a normal population with mean m and variance σ^2 . Then $(n-1)\hat{\Sigma}^2/\sigma^2$ is governed by the chi-squared density with $n-1$ degrees of freedom.

Equivalently, in the notation of Section 2, $\hat{\Sigma}^2$ has density $\frac{\sigma^2}{n-1} g_{(n-1)/2}(\frac{\sigma^2}{n-1} t; \frac{1}{2})$. By taking expectation the reader should observe that $\mathbb{E}(\hat{\Sigma}^2) = \sigma^2$: the sample variance is an *unbiased* estimator of the true variance vindicating the choice of scaling $n-1$ over n . For an alternative approach, see Problem XIV.6.

The reader may well feel now that a principled approach to constructing a confidence interval for the quality of the sample mean as an estimator of the mean is at hand: simply replace the unknown population variance σ^2 in the confidence interval by its unbiased estimate $\hat{\Sigma}^2$. This is philosophically sound but to make it practical we will wish to compute probabilities of the form $P\{| \hat{M} - m | \leq t\hat{\Sigma}/\sqrt{n}\}$ to determine our confidence that the estimate of the mean \hat{M} lies in a given interval $[m - t\hat{\Sigma}/\sqrt{n}, m + t\hat{\Sigma}/\sqrt{n}]$ of width $2t\hat{\Sigma}/\sqrt{n}$ centred at the unknown m . (The reader should note that the width of the interval now depends only on the *known* sample variance which is computed from the sample.) These probabilities are governed by the distribution of the *t-statistic* $T = \frac{\hat{M}-m}{\hat{\Sigma}/\sqrt{n}}$. To our dismay, however, a formidable *computational* objection to the procedure appears to present itself. While \hat{M} has a normal distribution and $\hat{\Sigma}^2$ has a chi-squared distribution, they are both determined by the *same* sample and it appears not at all trivial to sort out the dependence between them when we take a ratio of the quantities to form the t-statistic. A surprising discovery makes the calculation transparent.

THEOREM 2 *The sample mean and the sample variance of a random sample drawn from a normal population are independent random variables.*

PROOF: As before, by centring and scaling we may suppose that we are working with a standard normal population. Set $Y_j = Z_j - \hat{M}$ for $1 \leq j \leq n-1$, but set $Y_n = \hat{M}$. With $\mathbf{Y} = (Y_1, \dots, Y_{n-1}, Y_n)$, the linear transformation $\mathbf{X} \mapsto \mathbf{Y}$ is non-singular. [Indeed, if there existed constants c_1, \dots, c_{n-1}, c_n , not all zero, such that $c_1 Y_1 + \dots + c_{n-1} Y_{n-1} + c_n Y_n = 0$ then, writing $c = \frac{1}{n}(c_1 + \dots + c_{n-1} - c_n)$, we would have $(c_1 - c)X_1 + \dots + (c_{n-1} - c)X_{n-1} - cX_n = 0$ and the variables X_1, \dots, X_n would be degenerate.] The vector $\mathbf{Y} = (Y_1, \dots, Y_n)$ is hence centred and normal, and, in view of Theorem 4.4, its distribution is completely specified by the pairwise covariances of its components. Now, for $1 \leq j \leq n-1$,

$$\begin{aligned} Y_j Y_n &= \left(X_j - \sum_{k=1}^n \frac{X_k}{n} \right) \sum_{l=1}^n \frac{X_l}{n} = \frac{1}{n} \sum_{l=1}^n X_j X_l - \frac{1}{n^2} \sum_{k=1}^n \sum_{l=1}^n X_k X_l \\ &= \frac{1}{n} X_j^2 + \frac{1}{n} \sum_{l \neq j} X_j X_l - \frac{1}{n^2} \sum_{k=1}^n X_k^2 - \frac{1}{n^2} \sum_{k=1}^n \sum_{l \neq k} X_k X_l. \end{aligned}$$

As $E(X_k^2) = \text{Var}(X_k) = 1$ and $E(X_k X_l) = \text{Cov}(X_k, X_l) = 0$ if $k \neq l$, by additivity of the expectation integral, we may take expectations termwise on the right to obtain $\text{Cov}(Y_j, Y_n) = E(Y_j Y_n) = \frac{1}{n} - \frac{n}{n^2} = 0$ whence Y_j and Y_n are uncorrelated for each $1 \leq j \leq n-1$. By the corollary of the previous section it follows that the variable Y_n is independent of the system of variables (Y_1, \dots, Y_{n-1}) . To finish off the proof, we observe that $X_n = n\hat{M} - (X_1 + \dots + X_{n-1}) = \hat{M} - (Y_1 + \dots + Y_{n-1})$. It follows that $\hat{\Sigma}^2 = Y_1^2 + \dots + Y_{n-1}^2 + (Y_1 + \dots + Y_{n-1})^2$ is a function purely of Y_1, \dots, Y_{n-1} and is hence independent of $Y_n = \hat{M}$. ▶

The situation has clarified nicely. The t-statistic

$$T = \frac{\hat{M} - m}{\hat{\Sigma}/\sqrt{n}} = \frac{\hat{M} - m}{\sigma\sqrt{n}} \sqrt{\frac{(n-1)\hat{\Sigma}^2/\sigma^2}{n-1}} \quad (5.2)$$

is the ratio of a normally distributed variable to the square-root of an *independent* variable with the chi-squared distribution. It is cleanest to focus on this feature shorn of the notational complications introduced by the sample.

Suppose U, U_1, \dots, U_v are independent, standard normal random variables and let $V^2 = U_1^2 + \dots + U_v^2$. We are interested in the random variable $T = \frac{U}{V/\sqrt{v}}$.

The random variable V^2 is distributed as a chi-squared variable with v degrees of freedom. The scaled random variable $\frac{1}{v}V^2$ hence has the gamma density $\nu g_{\nu/2}(\nu x; 1/2) = g_{\nu/2}(x; \nu/2)$ so that its positive square-root $\sqrt{V^2/v}$

has density $2xg_{\nu/2}(x^2; \nu/2)$. As the random variables U and $\sqrt{V^2/\nu}$ are independent, *vide* (VII.10.2'), the ratio $T = \frac{U}{\sqrt{V/\nu}}$ has density

$$\begin{aligned}s_\nu(t) &= \int_0^\infty 2x^2 g_{\nu/2}(x^2; \nu/2) \phi(tx) dx \\ &= \frac{\nu^{\nu/2}}{2^{(\nu-1)/2} \sqrt{\pi} \Gamma(\frac{1}{2}\nu)} \int_0^\infty x^\nu \exp\left\{-\frac{\nu x^2}{2} \left(1 + \frac{t^2}{\nu}\right)\right\} dx.\end{aligned}$$

The change of variable $u = \frac{\nu x^2}{2} \left(1 + \frac{t^2}{\nu}\right)$ reduces the integrand to the form of a gamma density resulting in *Student's density with ν degrees of freedom*

$$s_\nu(t) = \frac{\Gamma(\frac{1}{2}(\nu+1))}{\sqrt{\nu\pi} \Gamma(\frac{1}{2}\nu)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

THEOREM 3 Let T be the t-statistic (5.2) that is engendered by a random sample of size n drawn from a normal population with mean m . Then T has Student's density with $n - 1$ degrees of freedom.

The reader who looks back at the procedure may very reasonably feel that replacing the variance σ^2 in the z-statistic by its estimate the sample variance $\hat{\Sigma}^2$ to form the t-statistic can only increase the uncertainty in how good the estimate of the mean really is. Her intuition is well-served. Student's density is strictly more variable than the standard normal density for any given number of degrees of freedom. In consequence, for any selected confidence level, the t-statistic leads to a wider confidence interval than that provided by the z-statistic. The uncertainty diminishes quickly with increasing sample size, however, and, as n becomes large, Student's density converges pointwise to the standard normal density. I will leave these assertions as exercises though the reader may find Figure 3 persuasive.

The t-statistic is due to W. S. Gosset who, while working for the Guinness brewery in Dublin, invented the test to study small samples for quality control in brewing. Worried about the leakage of trade secrets, Guinness had put in place a policy prohibiting its employees from publication and Gosset took to publishing his results under the pseudonym "Student" to circumvent the ban. His landmark paper on the t-statistic appeared in 1908.³ Here is E. S. Pearson on the influence of Gosset's work:⁴ "[Gosset's] investigation published in 1908 has done more than any other single paper to bring subjects within the range of statistical inquiry; as it stands it has provided an essential tool for the practical worker, while on the theoretical side it has proved to contain the seed of ideas which have since grown and multiplied in hundredfold." Gosset's

³Student, "The probable error of a mean", *Biometrika*, vol. 6, pp. 302–310, 1908.

⁴E. S. Pearson, "'Student' as statistician", *Biometrika*, vol. 30, pp. 210–250, 1939.

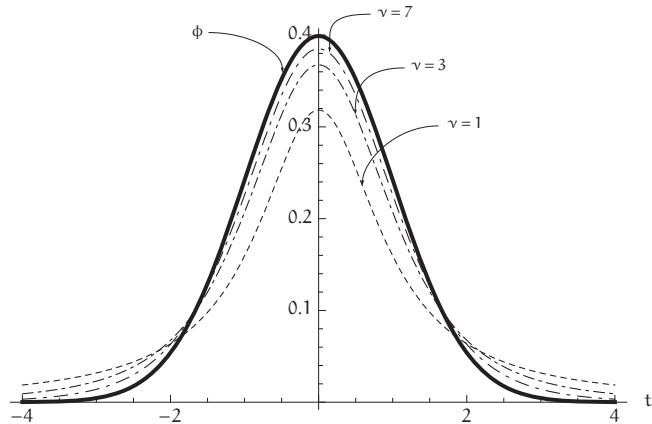


Figure 3: The graphs of Student's density with ν degrees of freedom for $\nu = 1, 3$, and 7 with the graph of the standard normal density superposed.

ideas on testing with small samples and the t-statistic have had and continue to have a huge influence in statistics and his pseudonym and original notation for the statistic have stuck.

The eminent statistician R. A. Fisher was aware of Gosset's work on small samples and proposed a related statistic for the analysis of variance.⁵ Fisher's ideas have been extremely influential, the all-purpose F-test a lineal descendant.

Suppose $X_1, \dots, X_\nu, X'_1, \dots, X'_\mu$ is a sequence of independent, standard normal random variables. Write $V^2 = X_1^2 + \dots + X_\nu^2$ and $V'^2 = X'_1^2 + \dots + X'_\mu^2$. The positive random variable $F = \frac{1}{\mu} V'^2 / \frac{1}{\nu} V^2$ is called the *F-statistic*.

Consider first the variable $W = V'^2/V^2$. Now V'^2 has density $g_{\nu/2}(\cdot; 1/2)$ while V^2 has density $g_{\nu/2}(\cdot; 1/2)$ and it is clear that V^2 and V'^2 are independent. Via another appeal to (VII.10.2'), the ratio $W = V'^2/V^2$ hence has density

$$\begin{aligned} w_{\mu, \nu}(t) &= \int_0^\infty x g_{\nu/2}(x; 1/2) g_{\mu/2}(tx; 1/2) dx \\ &= \frac{t^{\frac{1}{2}\mu-1}}{2^{\frac{1}{2}(\mu+\nu)} \Gamma(\frac{1}{2}\mu) \Gamma(\frac{1}{2}\nu)} \int_0^\infty x^{\frac{1}{2}(\mu+\nu)-1} e^{-\frac{1}{2}x(1+t)} dx. \end{aligned}$$

The change of variable $u = \frac{1}{2}x(1+t)$ results in the reduction of the integral to another gamma function yielding

$$w_{\mu, \nu}(t) = \frac{\Gamma(\frac{1}{2}(\mu+\nu))}{\Gamma(\frac{1}{2}\mu) \Gamma(\frac{1}{2}\nu)} \cdot \frac{t^{\frac{1}{2}\mu-1}}{(1+t)^{\frac{1}{2}(\mu+\nu)}}.$$

⁵R. A. Fisher, "The correlation between relatives on the supposition of Mendelian inheritance", *Philosophical Transactions of the Royal Society of Edinburgh*, vol. 52, pp. 399–433, 1918.

As the F-statistic is related to W by the change of scale $F = \frac{\nu}{\mu} W$, it follows that F has density

$$f_{\mu,\nu}(t) = \frac{\mu}{\nu} w\left(\frac{\mu}{\nu} t\right) = \frac{\mu \Gamma\left(\frac{1}{2}(\mu + \nu)\right)}{\nu \Gamma\left(\frac{\mu}{2}\right) \Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{\left(\frac{\mu t}{\nu}\right)^{\frac{1}{2}\mu-1}}{(1 + \frac{\mu t}{\nu})^{\frac{1}{2}(\mu+\nu)}}.$$

This is *Snedecor's density*, also known as the *F-density*. A variety of useful properties may be deduced from the functional form of the density. If, for instance, we write $T = T_\nu$ for Student's t-statistic to explicitly acknowledge the number of degrees of freedom and, likewise, $F = F_{\mu,\nu}$ for the F-statistic, it is clear that $F_{1,\nu}$ has the same distribution as T_ν^2 . The preceding observation says that $f_{1,\nu}(t) = \frac{1}{2\sqrt{t}} \{s_\nu(\sqrt{t}) + s_\nu(-\sqrt{t})\}$ as one may verify by direct calculation.

6 Echoes from Venus

The planet Venus is the brightest natural object in the skies after the sun and the moon. As our nearest planetary neighbour, it has fascinated humankind over the centuries, the discovery that it is blanketed by a thick cloud cover only adding to its allure, the fervid imagination populating the surface hidden from view with life forms and alien civilisations. It was only with the advent of the space age that the mysteries of Venus, obscured by its swirling clouds, could be peeled off, the earliest scientific examination of data from Venus dating to the 1960s.

While clouds are impermeable at optical wavelengths they are transparent at radio wavelengths leading to the idea that the Venusian surface could be mapped by radar. Earth-based radar facilities, however, faced formidable engineering challenges. The very great distances involved and limitations in the amount of power that could be transmitted meant that the radar returns were going to be so faint that extraordinary efforts would be needed to recover any information from them. Primary among the conditions requisite for the experiment was the need to ensure that the amount of electromagnetic noise from other sources be kept to an absolute minimum. An environment conducive to this need was found in a radio-quiet zone deep in the Mojave desert in California where the NASA/JPL Deep Space Instrumentation Facility at Goldstone is located.

In November 1962, ten kilowatts of radio frequency power were aimed at Venus from Goldstone and, after the 90 million kilometre round trip, the very faint radar echoes were recorded. As the signal was attenuated to a level of 10^{-21} watts, reception was at all possible only because of the quite extraordinary quality of the low-temperature ruby-cavity maser receiver that had been designed for the purpose. The signal consisted of the radio echo signature from the Venusian surface embedded in a sea of thermal noise contributed by the background radiation of the universe, other celestial electromagnetic sources,

and the receiver itself, and the task was to extract the faint traces of information from the signal.

The radar returns from Venus themselves carried information about surface features such as altitude, roughness and, by Doppler studies, rotation characteristics. As thermal noise has a normal characteristic this information was contained in the covariance between different points in the random waveform $X(t)$ that was received. The signal processor was however faced with the practical challenge that the faint information trace was imbedded in a very noisy waveform whence it was hard to extract reliable statistics from a single, or even a few, radar returns. As only the correlative structure of the random waveform $X(t)$ was informative, not the noise power itself, the simplest solution was to clip the waveform to one of two values—a process called *hard-limiting*—to create a new binary-valued waveform $Y(t) = \text{sgn } X(t)$ which takes value +1 if $X(t) \geq 0$ and value -1 if $X(t) < 0$. The procedure has much to recommend it:⁶ it eliminates the dominant, obfuscating noise power from the waveform and reduces computational complexity so drastically that a digital computer can make short work of computations. But the reader may feel that reducing the amplitude of the waveform to a single bit of precision is carrying matters a bit too far. How much precision is lost in the correlation estimate? Well, matters have been reduced to a purely mathematical question that we can set about modelling.

In a slight abuse of notation, write $X_1 = X(t_1)$ and $X_2 = X(t_2)$ for the values of the random waveform at two given points, and let $Y_1 = Y(t_1) = \text{sgn } X(t_1)$ and $Y_2 = Y(t_2) = \text{sgn } X(t_2)$ be the corresponding hard-limited values. With a proper normalisation of noise power we suppose that (X_1, X_2) forms a normal pair, each with zero mean and unit variance, and with covariance ρ reflecting their spatial dependency structure; that is to say, with notation as in (VII.5.3), the pair (X_1, X_2) has density given by the bivariate normal $\phi(x_1, x_2; \rho)$. Then Y_1 and Y_2 are symmetric ± 1 -valued random variables, $E(Y_1) = E(Y_2) = 0$. The product $Y_1 Y_2$ is another ± 1 -valued random variable. As $Y_1 Y_2 = +1$ if X_1 and X_2 have the same sign and $Y_1 Y_2 = -1$ if X_1 and X_2 are opposed in sign, we have

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= P\{Y_1 Y_2 = +1\} - P\{Y_1 Y_2 = -1\} = 2P\{Y_1 Y_2 = +1\} - 1 \\ &= 2P\{X_1 X_2 \geq 0\} - 1 = 4P\{X_1 \geq 0, X_2 \geq 0\} - 1. \end{aligned}$$

The expression simplifies to the form given in the final step on the right as, by symmetry, the pair (X_1, X_2) has equal probability of being in the first or in the third quadrant. The probability on the right is visualised most cleanly in vector notation. Writing $A = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for the covariance matrix of $\mathbf{X} = (X_1, X_2)$, we have

$$P\{X_1 \geq 0, X_2 \geq 0\} = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{\mathbb{I}} \exp\left(-\frac{1}{2}\mathbf{x}A^{-1}\mathbf{x}^T\right) d\mathbf{x}, \quad (6.1)$$

⁶R. M. Goldstein, "A technique for the measurement of the power spectra of very weak signals", *IRE Transactions on Space Electronics and Telemetry*, vol. 8, no. 2, pp. 170–173, 1962.

the integral on the right over the region $\mathbb{I} = \mathbb{R}^+ \times \mathbb{R}^+$ corresponding to the first quadrant of the plane. A change of variable of integration seems suggested, but which? A proper perspective makes the calculation virtually transparent.

By solving the eigenvalue equation $\mathbf{u}\Lambda = \lambda\mathbf{u}$ the reader may readily verify that the covariance matrix Λ has strictly positive eigenvalues $1 + \rho$ and $1 - \rho$ with corresponding eigenvectors $\mathbf{u}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\mathbf{u}_2 = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, both normalised to unit length. The eigensystem $\mathbf{u}_1, \mathbf{u}_2$ forms an orthonormal basis for \mathbb{R}^2 obtained simply by a 45° counter-clockwise rotation of the coordinate axes with the standard basis $\mathbf{e}_1 = (1, 0)$, $\mathbf{e}_2 = (0, 1)$. Writing $\mathbf{U} = (\mathbf{u}_1 \ \mathbf{u}_2)$ for the orthogonal matrix whose rows are the eigenvectors of Λ and $\Lambda = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$ for the diagonal matrix of eigenvalues, we may write the eigenvalue equation in the form $\Lambda = \mathbf{U}^T \Lambda \mathbf{U}$. The eigenvalues of Λ^{-1} are the reciprocals of the eigenvalues of Λ , the corresponding eigenvectors remaining unaltered, and so we also have the corresponding decomposition $\Lambda^{-1} = \mathbf{U}^T \Lambda^{-1} \mathbf{U}$ (see the discussion following Theorem 3.2).

As Λ is strictly positive definite, a suggestive further decomposition presents itself. We begin by defining the square-root of the diagonal matrix of eigenvalues Λ to be the diagonal matrix $\Lambda^{1/2} := \begin{pmatrix} \sqrt{1+\rho} & 0 \\ 0 & \sqrt{1-\rho} \end{pmatrix}$ whose diagonal entries are the positive square-roots of the corresponding diagonal elements of Λ . The definition is motivated by the fact that $\Lambda^{1/2} \cdot \Lambda^{1/2} = \Lambda$ and the familiar rule of exponents is preserved. The canonical diagonalisation of Λ now suggests how we may extend the definition to positive definite matrices: we define the *square-root* of Λ , denoted $\Lambda^{1/2}$, to be the matrix $\Lambda^{1/2} = \mathbf{U}^T \Lambda^{1/2} \mathbf{U}$. As

$$\Lambda^{1/2} \cdot \Lambda^{1/2} = \mathbf{U}^T \Lambda^{1/2} \mathbf{U} \mathbf{U}^T \Lambda^{1/2} \mathbf{U} = \mathbf{U}^T \Lambda^{1/2} \Lambda^{1/2} \mathbf{U} = \mathbf{U}^T \Lambda \mathbf{U} = \Lambda,$$

the rule of exponents is again preserved. It is clear that $\Lambda^{1/2}$ is symmetric and satisfies the eigenequation $\mathbf{U}\Lambda^{1/2} = \Lambda^{1/2}\mathbf{U}$ so that the square-root of Λ has the same eigenvectors as Λ with eigenvalues the square-root of the corresponding eigenvalues of Λ . It follows that $\det(\Lambda^{1/2}) = \det(\Lambda^{1/2}) = \sqrt{1 - \rho^2}$ and, *a fortiori*, $\Lambda^{1/2}$ is also strictly positive definite. By taking inverses, we also have

$$(\Lambda^{1/2})^{-1} = \mathbf{U}^T (\Lambda^{1/2})^{-1} \mathbf{U} = \mathbf{U}^T (\Lambda^{-1})^{1/2} \mathbf{U} = (\Lambda^{-1})^{1/2}.$$

If we set $\Lambda^{-1/2} := \begin{pmatrix} 1/\sqrt{1+\rho} & 0 \\ 0 & 1/\sqrt{1-\rho} \end{pmatrix}$ and define $\Lambda^{-1/2} := \mathbf{U}^T \Lambda^{-1/2} \mathbf{U}$, we may now identify $(\Lambda^{1/2})^{-1} = (\Lambda^{-1})^{1/2} = \Lambda^{-1/2}$. We verify the identities $\Lambda^{-1/2} \cdot \Lambda^{-1/2} = \Lambda^{-1}$ and $\Lambda^{1/2} \cdot \Lambda^{-1/2} = \mathbf{I}$, and the notation functions as it should.

With definitions in hand, we may express the density of \mathbf{X} in the form

$$\frac{\exp(-\frac{1}{2}\mathbf{x}\Lambda^{-1}\mathbf{x}^T)}{2\pi\sqrt{1-\rho^2}} = \frac{\exp(-\frac{1}{2}\mathbf{x}\Lambda^{-1/2}\Lambda^{-1/2}\mathbf{x}^T)}{2\pi\sqrt{1-\rho^2}} = \frac{\exp(-\frac{1}{2}\|\mathbf{x}\Lambda^{-1/2}\|^2)}{2\pi\sqrt{1-\rho^2}}.$$

The change of variable $\mathbf{x} \mapsto \mathbf{z} = \mathbf{x}\Lambda^{-1/2}$ simplifies the exponent to a sum of squares. Thus $\mathbf{Z} = (Z_1, Z_2) = \mathbf{X}\Lambda^{-1/2}$ has density $\det(\Lambda^{1/2})\phi(\mathbf{z}\Lambda^{1/2}; \rho) =$

$(2\pi)^{-1} e^{-(z_1^2 + z_2^2)/2}$ and, consequently, Z_1 and Z_2 are independent and normal, each with zero mean and unit variance. The probability on the right in (6.1) is hence equal to the probability that Z takes values in the region \mathbb{I}' corresponding to the image under $A^{-1/2}$ of the region \mathbb{I} .

Let $f_1 = e_1 A^{-1/2}$ and $f_2 = e_2 A^{-1/2}$ be the images under $A^{-1/2}$ of the standard basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$, respectively. As the positive half-axes determined by the unit standard basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$ form the boundaries of the region \mathbb{I} sweeping in the counter-clockwise direction from e_1 to e_2 , in the transformed coordinate system, the rays from the origin determined by the vectors f_1 and f_2 form the boundaries of the image \mathbb{I}' sweeping in a counter-clockwise direction from f_1 to f_2 . This is illustrated for the case $\rho = -0.75$ in Figure 4 where the transformation of the basis vectors $e_1 \mapsto f_1$, $e_2 \mapsto f_2$ is shown superimposed upon the circular level sets where the density of Z is constant. Passing from the Cartesian (Z_1, Z_2) system to polar coordinates (R, Θ) , we recall from Example VII.9.3 that Θ is uniformly distributed in the interval $(-\pi, \pi)$. It follows that $P\{Z \in \mathbb{I}'\} = \theta/(2\pi)$ where θ is the angle between f_1 and f_2 in the counter-clockwise direction.

It only remains now to determine θ . The law of cosines from elementary geometry tells us that $f_1 f_2^T = \|f_1\| \cdot \|f_2\| \cdot \cos(\theta)$. Computing the various inner products in turn, we have

$$\begin{aligned} f_1 f_2^T &= e_1 A^{-1} e_2^T = \frac{1}{1-\rho^2} (1 \quad 0) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{-\rho}{1-\rho^2}, \\ f_1 f_1^T &= e_1 A^{-1} e_1^T = \frac{1}{1-\rho^2} (1 \quad 0) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{1-\rho^2}, \\ f_2 f_2^T &= e_2 A^{-1} e_2^T = \frac{1}{1-\rho^2} (0 \quad 1) \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{1-\rho^2}. \end{aligned}$$

It follows that $\theta = \arccos(-\rho)$. Putting the pieces together, we obtain the unexpectedly simple formula

$$\text{Cov}(Y_1, Y_2) = \frac{4}{2\pi} \arccos(-\rho) - 1 = \frac{2}{\pi} \arcsin(\rho).$$

While thus far we have only encountered finite or denumerably infinite collections of random variables, the radar echo $X(t)$ represents a random waveform that is specified at a continuum of time instants. Such a waveform (or *random process*) is characterised by the specification of the distribution of the variables $(X(t_1), \dots, X(t_n))$ at every finite collection of time instants; these are the *finite-dimensional distributions* of the process. We say that the process

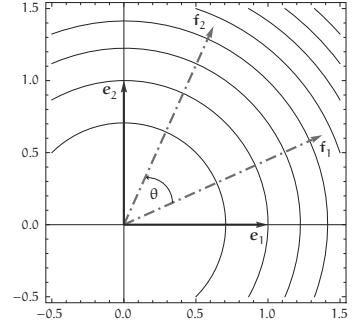


Figure 4: Coordinate transformation $A^{-1/2}$ for $\rho = -0.75$.

is *Gaussian* if all the finite-dimensional distributions are normal and that it is *stationary* if the finite-dimensional distributions are invariant with respect to shifts of the origin, that is to say, the processes $X(t)$ and $X(t + \tau)$ have the same finite-dimensional distributions for every τ . The radar echo $X(t)$ is now most naturally modelled as a sample function of a zero-mean, stationary Gaussian process. In view of Theorem 4.4, the finite-dimensional distributions of the process are completely specified by the correlations between $X(t)$ and $X(s)$ at any two points t and s , and as the process is stationary, these correlations depend only upon the absolute value of the differences $t - s$. Accordingly, all the finite-dimensional distributions are specified by an even function of one variable, $R(\tau) = E(X(t + \tau)X(t))$, called the *correlation function*. This leads to great simplifications. The value $R(0)$ is just the variance of $X(t)$ at each point t so the normalised process $X(t)/\sqrt{R(0)}$ is a zero-mean, stationary Gaussian process with unit variance and correlation function $R(\tau)/R(0)$. The hard-limited process $Y(t) = \text{sgn } X(t) = \text{sgn}(X(t)/\sqrt{R(0)})$ inherits stationarity from $X(t)$ and hence has all its finite-dimensional distributions determined by its correlation function $R_{\pm}(\tau) = E(Y(t + \tau)Y(t))$. Our surprising discovery that the correlation function of the original process is determined up to a scale factor by that of the hard-limited process may now be captured in a theorem and a slogan.

THEOREM $R_{\pm}(\tau) = \frac{2}{\pi} \arcsin\left(\frac{R(\tau)}{R(0)}\right)$.

SLOGAN *The hard-limited process carries all the essential statistical information of the originating stationary Gaussian process.*

This formulation was successfully applied by R. M. Goldstein in 1962 to show that Venus has a fluctuating roughness characteristic and, surprisingly, that Venus rotates retrograde with a period of about 250 days and with its axis nearly perpendicular to its orbit.⁷ Since the early days of earth-based radar, flybys and space probes, together with ever more sophisticated radar, have provided exquisite detail about the Venusian surface. But the huge costs associated with unmanned space exploration suggest that radar telemetry will continue to have a significant rôle to play in planetary exploration.

7 The strange case of independence via mixing

Suppose X_1 and X_2 are inputs to some physical system whose outputs are a linear form

$$\begin{aligned} Y_1 &= X_1 w_{11} + X_2 w_{21}, \\ Y_2 &= X_1 w_{12} + X_2 w_{22}. \end{aligned}$$

⁷R. M. Goldstein, "Venus characteristics by earth-based radar", *The Astronomical Journal*, vol. 69, no. 1, pp. 12–18, 1964.

We suppose to obviate trivialities that the matrix $W = [w_{jk}]$ is non-singular and further that W has at most one component that is zero. The outputs of such a system may be considered to be obtained by mixing the inputs and accordingly in such a context we say that a matrix of this form is a *mixing transformation*. Examples abound: to take three instances more or less at random, energy is transferred between modes in an optical fibre by mixing; communications from users of mobile telephones in the same region are additively tangled at the receiver; and medical imaging superposes images from several point sources.

While it is natural to model X_1 and X_2 as independent random variables if they originate from physically independent sources, it is intuitive to expect that mixing destroys independence. Indeed, we may suppose, if necessary by scaling X_1 and X_2 by $\det(W)$, that the mixing transformation has unit determinant. Writing f_1 and f_2 for the marginal densities of X_1 and X_2 , respectively, the induced density of the pair (Y_1, Y_2) is then given by $g(y_1, y_2) = f_1(y_1 w_{22} - y_2 w_{21}) f_2(-y_1 w_{12} + y_2 w_{11})$. It appears that the variables y_1 and y_2 are inextricably mixed in the expression for g and, if g_1 and g_2 represent the marginal densities of Y_1 and Y_2 , respectively, it is difficult to imagine how $g(y_1, y_2)$ may be separable into the product $g_1(y_1)g_2(y_2)$. The discussion at the end of Section 4 suggests how this may come about.

THEOREM 1 Suppose X_1 and X_2 are independent, standard normal random variables. If $W = \begin{pmatrix} \cos \vartheta & -\sin \vartheta \\ \sin \vartheta & \cos \vartheta \end{pmatrix}$ is a rotation matrix then Y_1 and Y_2 are independent and standard normal.

PROOF: Let $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ denote the identity matrix of order 2. Setting $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2) = \mathbf{X}W$, we have $\text{Cov}(\mathbf{Y}) = E(\mathbf{Y}^T \mathbf{Y}) = W^T E(\mathbf{X}^T \mathbf{X}) W = W^T I W = W^T W = I$, in a by now familiar manoeuvre by additivity of expectation. The covariance matrix of \mathbf{Y} is hence diagonal. The claimed result follows by the corollary of Section 4. ▶

The reader may wonder if normal densities are special in this regard. If this is so it must be because of the characteristic quadratic in the exponent of the density.

Suppose that X_1 and X_2 are independent with *continuous* marginal densities f_1 and f_2 , respectively, and suppose further that Y_1 and Y_2 are independent for some choice of mixing transformation W . Of course this is trivial if W is of the form $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ or $\begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}$ but this does not lead to an honest mixing which permits at most one zero among the components of W . We accordingly first consider the case where the matrix W has exactly one zero component. By scaling and relabelling the variables if necessary we may suppose that W is of the form $\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ where $a \neq 0$. Then the two-dimensional density induced by mixing satisfies $g_1(y_1)g_2(y_2) = f_1(x_1)f_2(x_2)$. But $Y_2 = X_2$ whence $g_2(y_2) = f_2(x_2)$ leading to $g_1(x_1 + ax_2) = f_1(x_1)$. But this is impossible and so, if Y_1 and Y_2 are to be independent then, if at all possible, it can only be if the mixing matrix W has no zero elements.

Suppose accordingly that $W = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ is an honest mixing matrix with no element identically zero. Then the two new coordinate axes perpendicular to the lines $x_1 w_{11} + x_2 w_{21} = 0$ and $x_1 w_{12} + x_2 w_{22} = 0$ cannot coincide with each other (as W is non-singular) nor can either of the new axes coincide with either of the old coordinate axes perpendicular to the lines $x_1 = 0$ and $x_2 = 0$ (as W can have no element identically zero). As before, we may suppose that W has unit determinant, if necessary by scaling both X_1 and X_2 by $\det(W)$. Again, the two-dimensional density induced by mixing is to satisfy

$g_1(y_1)g_2(y_2) = f_1(x_1)f_2(x_2)$ where now $y_1 = x_1w_{11} + x_2w_{21}$ and $y_2 = x_1w_{12} + x_2w_{22}$. Taking logarithms will enable us to isolate the exponents but we have to guard against the possibility that one or more of the densities may vanish in some interval.

To see that this cannot be the case suppose Ω is any maximal connected region in the plane in which the densities $f = f_1f_2$ and $g = g_1g_2$ are strictly positive. In other words, Ω is a *component*. If Ω is not all of \mathbb{R}^2 then, by definition, $f = g = 0$ on the boundary $\partial\Omega$ of Ω . Now if $f = 0$ then $f_1 = 0$ or $f_2 = 0$ and $\partial\Omega$ must consist of lines $x_1 = \text{constant}$ or $x_2 = \text{constant}$ perpendicular to the coordinate axes. On the other hand, if $g = 0$ then $g_1 = 0$ or $g_2 = 0$ and $\partial\Omega$ must consist of lines $y_1 = \text{constant}$ or $y_2 = \text{constant}$ perpendicular to the new (transformed) coordinate axes. But both conditions cannot hold simultaneously (else we would have the contradiction that $f = g = 0$ at points in the interior of Ω). It must hence be the case that none of f_1, f_2, g_1 , or g_2 vanish anywhere.

We may now safely take logarithms of these densities and so let $\gamma_1 = \log g_1$, $\gamma_2 = \log g_2$, $\varphi_1 = \log f_1$, and $\varphi_2 = \log f_2$. Then

$$\gamma_1(y_1) + \gamma_2(y_2) = \varphi_1(x_1) + \varphi_2(x_2). \quad (7.1)$$

To verify that each of these functions is a quadratic it will suffice to show that their second derivatives are constant. But here we run into the objection that while these functions are given to be continuous, we do not know *a priori* whether they are differentiable, much less twice differentiable. Well, no matter, differences will do as well, these not requiring the fine structure of differentiability. Suppose h_1 and h_2 are fixed and strictly positive. If $u(x_1, x_2)$ is any function of two variables we introduce the *centred second difference operator* with the nonce notation⁸

$$\begin{aligned} \Delta^2 u(x_1, x_2) &= u(x_1 + h_1, x_2 + h_2) - u(x_1 + h_1, x_2 - h_2) \\ &\quad - u(x_1 - h_1, x_2 + h_2) + u(x_1 - h_1, x_2 - h_2). \end{aligned}$$

By identifying u with the two sides of (7.1) and applying the linear operator Δ^2 to both sides, we obtain $\Delta^2\gamma_1 + \Delta^2\gamma_2 = \Delta^2\varphi_1 + \Delta^2\varphi_2$ where, in a slight abuse of notation, we identify the four functions as implicit functions of the two variables x_1 and x_2 . Of course, the function $\varphi_1(x_1)$ is invariant with respect to x_2 and, likewise, the function $\varphi_2(x_2)$ is invariant with respect to x_1 , and it is hence easy to verify that $\Delta^2\varphi_1 = \Delta^2\varphi_2 = 0$. It follows that $\Delta^2\gamma_1 = -\Delta^2\gamma_2$. The expression on the left is a function of $y_1 = x_1w_{11} + x_2w_{21}$, while the expression on the right is a function of the linearly independent variable $y_2 = x_1w_{12} + x_2w_{22}$. We may hence vary x_1 and x_2 keeping, say, y_2 fixed and allowing y_1 to vary over all values, and so it follows that $\Delta^2\gamma_1$ and $\Delta^2\gamma_2$ must each be independent of the coordinate variables x_1 and x_2 and are both hence determined solely by the choice of parameters h_1 and h_2 .

Now, by setting $t_1 = h_1w_{11} + h_2w_{21}$ and $t_2 = h_1w_{11} - h_2w_{21}$ we may write $\Delta^2\gamma_1$ in the form $\gamma_1(y_1 + t_1) - \gamma_1(y_1 + t_2) - \gamma_1(y_1 - t_2) + \gamma_1(y_1 - t_1)$. We may

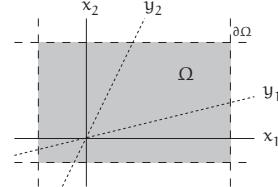


Figure 5: A component.

⁸The operator reappears in Sections XVI.5 and XIX.7 where it is convenient to redefine it in a one-sided formulation.

simplify matters further by selecting h_1 and h_2 such that $t_1 = t$ and $t_2 = 0$ whence $\Delta^2\gamma_1 = \gamma_1(y_1 + t) - 2\gamma_1(y_1) + \gamma_1(y_1 - t)$. On the other hand, $\Delta^2\gamma_1$ is determined solely by the choice of h_1 and h_2 which are completely specified in terms of t . It follows that $\Delta^2\gamma_1$ is a function only of the parameter t at our discretion. Accordingly,

$$\gamma_1(y_1 + t) - 2\gamma_1(y_1) + \gamma_1(y_1 - t) = v(t) \quad (7.2)$$

for some continuous function v .

We have simplified matters to a consideration of second differences of the exponent of one of the densities in question and, as we are entertaining the suspicion that the density is normal, we should promptly check to see how the second difference of a quadratic behaves. Accordingly, let $\tilde{\gamma}(y_1) = ay_1^2 + by_1 + c$ be any quadratic with coefficients to be specified. Then $\tilde{\gamma}_1(y_1 + t) - 2\tilde{\gamma}_1(y_1) + \tilde{\gamma}_1(y_1 - t) = 2at^2$ and a generic quadratic satisfies an equation of the form (7.2) with a different choice of continuous function $\tilde{v}(t) = 2at^2$ on the right. The difference $\tilde{\gamma}_1(y_1) = \gamma_1(y_1) - \tilde{\gamma}_1(y_1)$ hence also satisfies an equation of the form

$$\tilde{\gamma}_1(y_1 + t) - 2\tilde{\gamma}_1(y_1) + \tilde{\gamma}_1(y_1 - t) = \tilde{v}(t) \quad (7.2')$$

for yet another continuous function $\tilde{v}(t) = v(t) - \tilde{v}(t)$. We wish to test whether $\tilde{\gamma}_1(y_1)$ is, in fact, identically zero for a suitable choice of quadratic $\tilde{\gamma}_1(y_1)$. This appears difficult to manage on the face of it but, certainly, by choosing the constants a , b , and c at our discretion we may force $\tilde{\gamma}_1(y_1) = \gamma_1(y_1) - (ay_1^2 + by_1 + c)$ to be zero at any three specified points y'_1 , y''_1 , and y'''_1 . But any continuous function $\tilde{\gamma}_1$ that is zero at three points must have both a maximum and a minimum. (This is a consequence of Bolzano's theorem which the reader may recall asserts that a continuous function which takes values of opposite sign at the two endpoints of an interval must have a zero within the interval.) Now, in a small enough neighbourhood of a minimum of $\tilde{\gamma}_1$ we must have $\tilde{\gamma}_1(y_1 + t) - 2\tilde{\gamma}_1(y_1) + \tilde{\gamma}_1(y_1 - t) \geq 0$ and so $\tilde{v}(t) \geq 0$ in a neighbourhood of the origin. Contrariwise, in a small enough neighbourhood of any maximum of $\tilde{\gamma}_1$ we must have $\tilde{\gamma}_1(y_1 + t) - 2\tilde{\gamma}_1(y_1) + \tilde{\gamma}_1(y_1 - t) \leq 0$ and so $\tilde{v}(t) \leq 0$ in a neighbourhood of the origin. This forces $\tilde{v}(t)$ to be identically zero in a vicinity of the origin. But this means that $\tilde{\gamma}_1$ is identically zero so that γ_1 must be a quadratic. A density with a quadratic exponent must be normal and so g_1 is normal. A similar argument shows that g_2 is normal and, by reversing the rôles of the pairs (x_1, x_2) and (y_1, y_2) , f_1 and f_2 are also normal—as the reader may have suspected, Theorem 1 indeed captures a special attribute of the normal.

THEOREM 2 Suppose X_1 and X_2 are independent with continuous marginal densities and suppose further that $Y_1 = X_1 w_{11} + X_2 w_{21}$ and $Y_2 = X_1 w_{12} + X_2 w_{22}$ are variables obtained by an honest mixing transformation. If Y_1 and Y_2 are independent then all four variables are normal.

This theorem was generalised by degrees until a quite general form was proved independently by V. P. Skitovich and G. Darmois.⁹ Their theorem: let X_1, \dots, X_n be independent random variables and let $Y_1 = \sum_{j=1}^n a_j X_j$ and $Y_2 = \sum_{j=1}^n b_j X_j$ where no coefficient is identically zero. If Y_1 and Y_2 are independent then the variables X_1, \dots, X_n are normal. In

⁹V. P. Skitovitch, "On a property of the normal distribution", *Doklady Akademii Nauk SSSR*, vol. 89, pp. 217–219, 1953. G. Darmois, "Analyse générale des liaisons stochastiques", *Rev. Inst. Internationale Statist.*, vol. 21, pp. 2–8, 1953.

particular, the theorem applies also to situations where the constituent variables may not have a density (the so-called singular cases). My excuse for not presenting it here in full generality is that, while the general case builds upon the particular insight on the quadratic exploited in our proof, a formidable apparatus from analytic function theory is needed to deal with the Fourier transforms of distributions used in the proof.

Postscript: I must confess that I had always thought that the Skitovich–Darmois theorem was illustrative once more of a characteristic feature of the normal but had no particular utility otherwise. To my delight researchers in the 1990s unearthed an unlooked for and very pretty application.¹⁰ This goes to show that even obscure theorems can have their uses.

Suppose X_1 and X_2 denote symbols emitted by independent sources which we assume to be possessed of continuous marginal densities and let $\mathbf{X} = (X_1, X_2)$ denote the emitted symbol vector. In typical applications X_1 and X_2 are not normal though it will suffice for our purposes if we assume that at least one is not.¹¹ If the setting is one of communication over a shared medium such as a wireless channel then a tower or base station in contact with both sources obtains a mixed version of the two symbols. If there are two receiving antennae then two mixed versions of the symbols are received, the received pair $\mathbf{Y} = (Y_1, Y_2)$ satisfying the mixing equation $\mathbf{Y} = \mathbf{X}\mathbf{W}$ where the mixing matrix $\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ is characteristic of the intervening medium or *channel*. The same situation occurs when two speakers in a common area speak simultaneously, their utterances picked up by two microphones; or, when two scatterers in an ultrasound image are imaged in two distinct views.

If the mixing matrix is known then, computational issues aside, one can in principle recover the pair X_1 and X_2 by inverting the mixing equation to obtain $\mathbf{X} = \mathbf{Y}\mathbf{W}^{-1}$. The difficulty, however, is that the mixing transformation (in the jargon, the *channel matrix*) is not known and much effort is expended in estimating its components, typically by transmitting known pilot signals. Theorem 2 suggests an intriguing alternative.

With \mathbf{V} any square matrix of order 2, form the random vector $\mathbf{Z} = \mathbf{Y}\mathbf{V}$. If $\mathbf{Z} = (Z_1, Z_2)$ turns out to have independent components, what then can be concluded? As $\mathbf{Z} = \mathbf{X}\mathbf{W}\mathbf{V}$ is a linear transformation of the inputs it is clear that $\mathbf{W}\mathbf{V}$ must be non-singular else the components of \mathbf{Z} would be deterministically linearly related. In view of Theorem 2 of the previous section, however, a non-normal vector \mathbf{X} with independent components cannot engender a vector \mathbf{Z} with independent components via a mixing transformation. It must follow that $\mathbf{W}\mathbf{V}$ is non-singular but not a mixing matrix. But then $\mathbf{W}\mathbf{V}$ must have exactly one non-zero element in each row and in each column or, in other words, for some non-zero a and b , $\mathbf{W}\mathbf{V}$ is of the form $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$ or $\begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}$ obtained by multiplying a diagonal matrix and a permutation matrix. It follows that Z_1 and Z_2 are copies, possibly scaled and permuted, of X_1 and X_2 .

Thus, if the received vector $\mathbf{Y} = (Y_1, Y_2)$ is passed through a subsequent mixing matrix \mathbf{V} so chosen that $\mathbf{Z} = \mathbf{Y}\mathbf{V}$ has independent components then the components of \mathbf{Z} are copies of the input symbols up to scales and a permutation. While the channel matrix \mathbf{W} is in general unknown and outside our control, the selection of the matrix \mathbf{V} is entirely within our

¹⁰C. Jutten and J. Herault, “Blind separation of sources. Part I: An adaptive algorithm based on neuromimetic architecture”, *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

¹¹If both X_1 and X_2 are normal then it will suffice to replace, say, X_1 by a one-to-one, continuous, non-linear transform, $X_1 \leftarrow \alpha(X_1)$, which does not change the problem in essentials.

jurisdiction; the Skitovich–Darmois theorem suggests that a principled approach would be to select V to separate the sources statistically, effectively unravelling the mixing performed by the channel. On reflection the reader may find it remarkable that this can be accomplished without any knowledge of the channel matrix solely by a consideration of the statistics of the output symbols. This procedure is hence called *blind source separation* or *blind channel equalisation*.

There are of course many practical hurdles. How does one go about testing for independence at the output? One measure is suggested by the fact that independent variables are uncorrelated. An *ad hoc* procedure may then be crafted to iteratively adapt the matrix V to decorrelate the output symbols. In practice, of course, one would have to use a random sample of transmit–receive pairs to form an estimate of the correlation. While this is a relatively straightforward procedure it is very crude as uncorrelatedness is a poor indicator of independence as we have non-normal variables at hand. A more principled approach is suggested by Problem XIV.44.

8 A continuous, nowhere differentiable function

In 1872, to the great consternation of mathematicians of the time, Karl Weierstrass constructed a continuous function that was nowhere differentiable. It is difficult to think of such a function directly so let us edge up to it. Consider the sequence of functions $h_n(t) = \sum_{k=1}^n 2^{-k/2} \cos 2^{k+1}\pi t$ on the unit interval $[0, 1]$. These functions are a refinement of the original Weierstrass construction due to the British mathematician G. H. Hardy. Now each of the h_n is well-behaved and as the series coefficients provide damping by an exponentially decreasing factor we anticipate that the sum should be nicely convergent. Trouble rears its head in the frequencies of the cosines in the sum. Though they are getting damped ferociously, as k increases the summands are wiggling faster and faster and we can imagine the incorporation of an infinite sequence of more and more rapid, infinitesimal wiggles into the function h_n as n becomes large.

Some pictorial evidence is worth any amount of algebraic posturing here and accordingly I have shown the function $h_{100}(t)$ graphed at varying time scales in Figure 6. Going clockwise from the upper left corner, we focus on successively shorter intervals centred at 0.25. The reader should observe how the essentially jagged nature of the curve is preserved however magnified our picture is. As we increase n the function changes rapidly on smaller and smaller time scales and in the limit it is not hard to imagine, with a small leap of faith, that we have a continuous function which appears kinky (microscopically) everywhere. Indeed, Weierstrass's function is the earliest analytical example of what we now call a *fractal*, a term from the Latin *fractus* meaning “broken”, that was coined and popularised in 1975 by Benoît Mandelbrot to describe curves and objects that exhibit this kind of self-similarity property over all scales.

In view of the graphical evidence, it does not place too large a burden on imagination to entertain the idea that the sequence of functions h_n

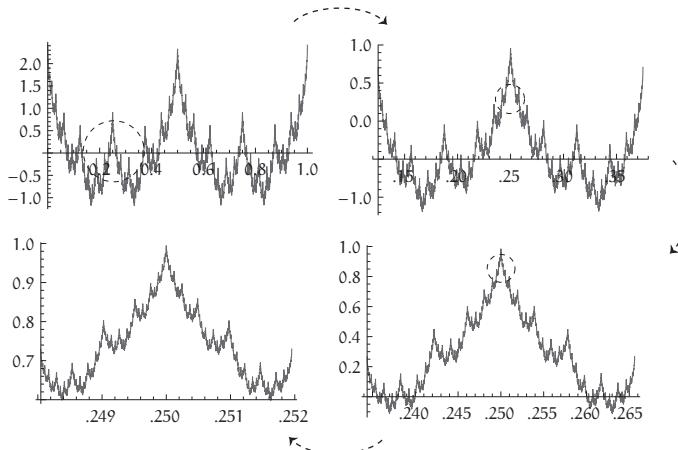


Figure 6: Fractal time scales for Hardy's continuous, nowhere differentiable function.

converges to a limiting function h which has the requisite property. As indeed it does: *the sequence $h_n(t)$ converges pointwise to an honest function $h(t) = \sum_{k=1}^{\infty} 2^{-k/2} \cos 2^{k+1} \pi t$ which is continuous everywhere on the unit interval $[0, 1]$ but, remarkably, differentiable nowhere on it!* The reader who has not seen this before may well be inclined to take up the White Queen's salutary habit of believing six impossible things before breakfast. While the proof is not so hard that it cannot be presented here it will take us a little too far afield and I shall be happy with the pictorial evidence. (Problem 26 provides another, somewhat more tractable, example for the curious reader.)

Weierstrass opened the floodgates and a flood of curves which have no derivative poured out. And not just in one dimension. Following on the heels of Weierstrass's discovery Giuseppe Peano discovered in 1890 that every point in space could be covered by a single, very wiggly, curve. Other space-filling curves were soon discovered by Hilbert, Moore, Lebesgue, and Sierpiński.

The extent of the affront to the mathematical intuition of the time that was caused by Weierstrass's construction and the seemingly endless nowhere differentiable offspring that it engendered may be gauged by the reactions of some of his contemporaries. Hermite called them "This dreadful plague of continuous, nowhere differentiable functions"; and Poincaré wrote "We have seen a rabble of functions whose only job, it seems, is to look as little as possible like decent and useful functions". In retrospect, the negative press was somewhat overblown. The study of these nowhere differentiable curves was to lead to fruitful new areas of inquiry and new directions in mathematical analysis. Far from being pathological, to add to our native disquiet, in a certain formal sense "most" curves are of this character as we shall see in Section 11.

Nature appears to leave a fractal footprint in divers places: snowflakes,

clouds, raindrop patterns, defensive colouration and camouflage, river beds, blood vessels, patterns of fracture and breakdown in materials, the humble broccoli and cauliflower, mountain ranges, and the British coastline all are approximately fractal.¹² But fractals are not merely a kind of mathematical conceit introduced as a model or description of nature. The nowhere differentiable curves ushered in by Weierstrass not only had a huge mathematical impact but had fecund application in art, science, engineering, and mathematical finance.

In the arts, fractal designs have appeared in Moorish art and architecture in the mediaeval world and, more recently, in the work of the American painter Jackson Pollock. In cutting-edge research, engineers are using fractal designs for antennas in micro-arrays, dense optical storage in volume holograms using fractal arrangements, signal and image compression via fractals, and in the design of efficient computer algorithms using space-filling curves. This little digression is intended to convince the reader of the wealth of application and ideas ushered in by Weierstrass's discovery and perhaps induce her to read more on it. Its connection to the normal distribution is our next subject.

9 Brownian motion, from phenomena to models

In 1827 the Scottish botanist Robert Brown observed in the study of pollen grains in water that the motion of the individual grains was extremely jittery. The appellation "Brownian motion" was born. Brownian motion as a model of physical phenomena was legitimised by Albert Einstein in one of his legendary papers of 1905 (the *annus mirabilis*) in which he examined the possibility that molecular motions could be explained by Brownian motion. Einstein's calculations formed the basis of J. B. Perrin's intricate experiments of 1908 to estimate Avogadro's number—the number of molecules in a mole of gas; current estimates put it at about 6×10^{23} . These experiments cemented Brownian motion as an important phenomenological tool in the hand of the natural scientist.

In Perrin's study of the trajectories of individual particles he described them as follows. "The trajectories are confused and complicated so often and so rapidly that it is impossible to follow them The apparent mean speed of a grain during a given time varies in *the wildest way* in magnitude and direction, and does not tend to a limit as the time taken for observation decreases It is impossible to fix a tangent, even approximately, and we are thus reminded of the continuous, non-differentiable functions of the mathematicians" This passage quoting Perrin is taken from Paley and Wiener.¹³ The thread linking Brown to Weierstrass lies in the informed guess that a mathematical model of

¹²B. Mandelbrot, "How long is the coast of Britain? Statistical self-similarity and fractional dimension", *Science*, New Series, vol. 156, no. 3775, pp. 636–638, 1967.

¹³R. E. A. C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain*, p. 127. Coll. Publ. Amer. Math. Soc., Providence, RI, 1934.

the motion of an individual Brownian grain (a “sample path” in modern parlance) could be a candidate for a continuous nowhere differentiable function.

Prior to Einstein’s work, Brownian motion had also made a rather peculiar appearance in 1900 in the doctoral thesis of Louis Bachelier at the Sorbonne with the provocative title *Théorie de la Spéculation*. Bachelier was a student of Poincaré and, rather ironically in light of Poincaré’s feelings about the Weierstrass function, had attempted to use Brownian motion to model the behaviour of stock and option markets. In the early part of the twentieth century then Brownian motion had gained importance in the natural sciences ... and a somewhat outré reference in financial markets.

While Einstein was aware of Bachelier’s work and, as we have seen, used Brownian motion models to great effect in particle physics, Bachelier’s ideas on finance on the other hand were to languish in obscurity for a half-century. The story picked up in the 1950s when a query from James Savage about Bachelier’s work began to make the rounds. Savage’s question came to the attention of the formidable economist Paul A. Samuelson who was working on options pricing at that time. In a story with a romantic cast to it, Samuelson rediscovered Bachelier’s forgotten thesis and was enraptured by the ideas it contained—the ugly duckling was about to blossom into an elegant swan. Samuelson refined and extended Bachelier’s work and in his seminal 1965 paper *Rational Theory of Warrant Pricing* argued that Brownian motion could be used to model stock prices. The stochastic differential equation that would become the key assumption in the magisterial Black–Scholes options pricing formula made its appearance in this paper.¹⁴ As Samuelson was an *éminence grise* in American economics, his paper had great influence. The Brownian motion model for stock markets led in turn to the famous formula for options pricing discovered by Fischer Black and Myron S. Scholes in 1973 and bearing their name. The derivation of the Black–Scholes formula is technical and will take us a bit too far afield, but the problem itself is easy to describe. In Scholes’s words, “The problem it’s trying to solve is to determine the value of the right, but not the obligation, to buy a particular asset at a specified price, within or at the end of a specified time period”. And under certain conditions that is what the Black–Scholes formula does through the construction of a self-financing, dynamic portfolio strategy. The door had been opened to the holy grail of risk-free portfolios. With significant contributions from Samuelson’s student Robert C. Merton, this led eventually to the Nobel Prize in economics for Merton and Scholes, Black having sadly passed away in 1995 just two years before the Nobel was awarded for the discovery of the formula that he had been instrumental in bringing to the light of day. Weierstrass’s scions had borne noble—and Nobel—fruit.

¹⁴When it was put to him that many people were unaware that the equation that was key to the Black–Scholes formula appeared in his 1965 paper, Samuelson is reported to have wryly said, “Yes, I had the equation, but they had the formula”.

So, continuous sample paths without derivatives moved from biology to mathematics, then to physics, and finally to economics and high finance.¹⁵ What are the features of these Brownian random trajectories described so vividly by Perrin that may be abstracted into a mathematical model of a random Brownian motion waveform $B(t)$ originating at $t = 0$?

Einstein's study of the statistical behaviour of Brownian particles suggests that the motion is engendered by a large number of (independent) molecular forces so that in view of the central limit theorem it may be reasonable to suppose that, at each t , the value $B(t)$ is a normal variate and more that, at any finite collection of times t_1, \dots, t_n , the collection $B(t_1), \dots, B(t_n)$ is jointly normal. If there is no drift then $B(t)$ should have zero mean and as observation suggests an increasing spread proportional to the square-root of t we should set its variance proportional to t . Moreover, in view of an idealised independence in the molecular interactions it is natural to posit that the process has *independent increments*, that is to say, for any $t_1 < t_2 < \dots < t_n$ the random variables $B(t_1), B(t_2) - B(t_1), \dots, B(t_n) - B(t_{n-1})$ are independent. This requirement completely captures the covariance structure of the process: indeed, if $s < t$, we may write $B(t) = B(s) + [B(t) - B(s)]$, and so $\text{Cov}(B(s), B(t)) = E[B(s)^2] + E[B(s)(B(t) - B(s))] = \text{Var}(B(s))$ in view of the assumed lack of drift and the independence of the increments. Finally, physical intuition suggests that trajectories should be unbroken, a point of view quite in accordance with Perrin's observations, and so we should require that $B(t)$ be continuous, at least with probability one. From these considerations arises a

DEFINITION A standard Brownian motion $B(t)$ on the unit interval $[0, 1]$ begins at the origin and satisfies the following conditions:

- ① With probability one $B(t)$ is continuous.
- ② For each integer $n \geq 1$ and collection of points $0 < t_1 < \dots < t_n < 1$, the vector $(B(t_1), \dots, B(t_n))$ has a normal distribution centred at the origin and specified by the covariances $\text{Cov}(B(t_j), B(t_k)) = \min\{t_j, t_k\}$.

The second condition is equivalent to the statement that the process $B(t)$ has independent, normal increments with $B(t) - B(s) \sim \mathcal{N}(0, t - s)$ for each pair $s < t$, as the reader should verify. Accordingly, Brownian motion is sometimes defined in terms of the independent increment property.

¹⁵ A political postscript: Merton and Scholes were directors in the infamous hedge fund *Long Term Capital Management*. From 1994 to 1997 the fund was the darling of Wall Street and made enormous profits for its wealthy investors. And then in 1998 the perfect financial storm hit. The firm ended up losing \$4.6 billion in less than four months because of a confluence of risk factors exacerbated by poor management. The fund was bailed out by the Federal Reserve Bank which was worried about dire consequences for world financial markets if the fund were allowed to fail, a move by the federal reserve chairman Alan Greenspan that is still hotly debated in some quarters. The slogan? Hedge funds carry inherent risk even with good management.

The reader may be pardoned for feeling that this is a rather formidable set of conditions to place upon the unsuspecting Brownian particle which may, understandably, begin to feel rather oppressed. Indeed there is a significant danger here that the definition is so restrictive in fact that it is completely vacuous. In other words, is Brownian motion merely a convenient mathematical fiction modelling aspects of reality or is it a real mathematical object?

The answer was provided by Norbert Wiener's very difficult investigations beginning in 1923. (Perhaps the only person who really understood Wiener's early work was Paul Lévy who had begun to think along similar lines.) While Wiener initially approached Brownian motion by first constructing a measure (the Wiener measure) on the space of trajectories and then imbedding a process, interactions with Paley and Zygmund convinced him of the virtues of an orthogonal series approach. This is the line we will follow.

Let $Z_0, Z_1, Z'_1, Z_2, Z'_2, \dots$ be a sequence of independent, standard normal random variables. (The reader who is a stickler for detail, or worse, is of a suspicious disposition, and objects that we haven't shown that such a sequence can be constructed will find questions of existence tackled in Section XII.9.) Wiener's theorem: the trigonometric series with random coefficients defined by

$$B(t) = Z_0 t + \sum_{n=1}^{\infty} \frac{\sqrt{2}}{2\pi n} [Z_n(1 - \cos 2\pi n t) + Z'_n \sin 2\pi n t] \quad (9.1)$$

converges with probability one to a standard Brownian motion on the unit interval. Truncations $B_N(t)$ of the sum to N terms suggest a practical method for the construction of Brownian motion. The reader may find Figure 7, which shows one such trajectory for a sum truncated to 5000 terms, persuasive. Wiener's inves-

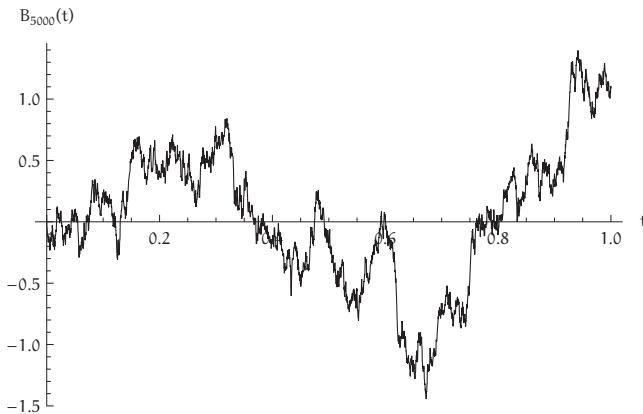


Figure 7: The truncated Wiener process.

tigations mathematically legitimised Brownian motion and put its burgeoning

applications on a sound logical footing. In recognition of his pioneering work in many quarters Brownian motion is hence referred to as the *Wiener process*.

The main difficulty in the proof of Wiener's theorem is in showing that the random trigonometric series given in (9.1) converges and Wiener's demanding original proofs required a substantial amount of mathematical dexterity. While it was entirely understandable given the period when Wiener had begun to consider these questions that he opted for a Fourier series with random coefficients, many of the technical difficulties vanish if an orthonormal basis of functions is chosen carefully. In view of the huge importance of Brownian motion I feel compelled to give an honest account of its construction in the concluding sections of this chapter. This material is not used in the sequel.



10 The Haar system, a curious identity

The theory of wavelets saw a boom in the latter part of the twentieth century as an alternative to traditional Fourier analysis. In many ways, these systems of functions are much more intuitive than the ubiquitous trigonometric functions of the Fourier system that the reader will recall seeing in divers contexts. A natural domain where wavelet models excel is in multi-spectral analysis in applications where data arise in multiple time and frequency scales. A key feature explaining the success of wavelets in such applications is the sharp localisation of these systems. This is the property that we will exploit in the construction of Brownian motion.

The natural domain of operation of wavelets is the space of square-integrable functions. It will suffice for our purposes to consider the real vector space $L^2[0, 1]$ of functions $f: [0, 1] \rightarrow \mathbb{R}$ for which the integral $\int_0^1 f(x)^2 dx$ exists and is convergent. While this space has many beauties to recommend it, the only result that we will need from L^2 theory is Parseval's equation which, in the language of physics, is a statement of energy conservation. The reader new to this space should glance at Section XXI.3 in the Appendix before reading on.

The archetypal wavelets are those comprising the Haar system. We begin with the function $h_1(x)$ which takes value $+1$ for $0 \leq x < 1/2$ and value -1 for $1/2 \leq x < 1$. By allowing j and k to range over integer values $j \geq 0$ and $0 \leq k \leq 2^j - 1$, the terms $n = 2^j + k$ range over all positive integers and we define $h_n(x) = h_{2^j+k}(x) = 2^{j/2} h_1(2^j(x - k2^{-j}))$. We finish up the specification by setting h_0 to be the indicator function for the unit interval, $h_0(x) = 1$ for $0 \leq x < 1$.

A picture is worth any amount of algebraic posturing here and Figure 8 illustrates the essential nature of this system of functions. The "father" and "mother" functions for this system are h_0 and h_1 , each with support in the unit interval. The rest of the functions, the "children", are engendered from h_1 by successive scales of axis by the factor $1/2$ and shifts of origin. Thus, h_2 and h_3 have support in non-overlapping intervals of width $1/2$, h_4, h_5, h_6 , and h_7 have support in non-overlapping intervals of width $1/4$, and so on. We may visualise these functions in rows starting with h_1 in row number 0, the j th row consisting of a serried rank of 2^j successive functions marching from left to right with support in successive dyadic intervals of width 2^{-j} . For natural reasons these functions are called *wavelets*, the localised waves rippling from left to right before

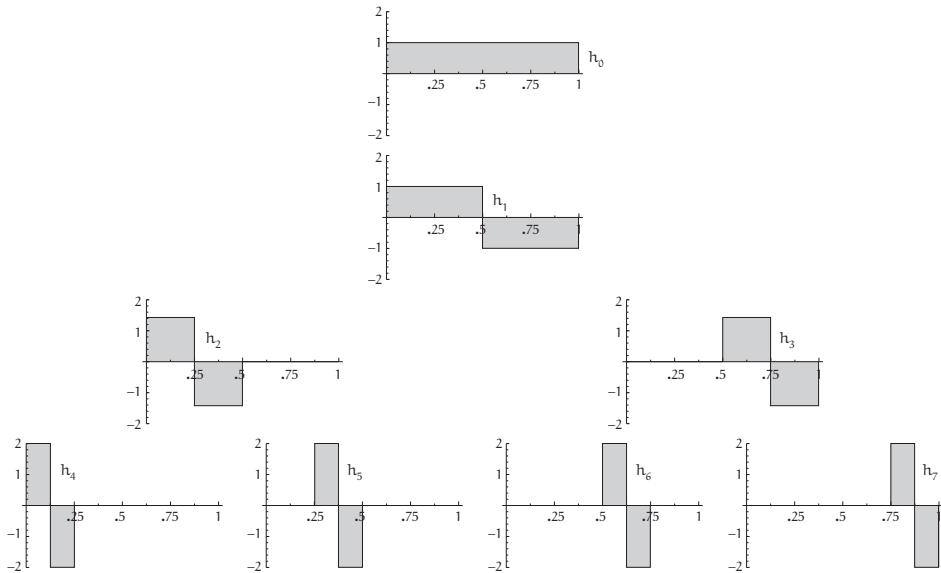


Figure 8: The Haar wavelet system.

returning, becoming narrower and higher, and proceeding from left to right again.

The Haar functions are most naturally viewed as living in the real vector space $L^2[0, 1]$ of square-integrable functions on the unit interval. This space is equipped with a notion of projection through the natural inner product $\langle f, g \rangle = \int_0^1 f(x)g(x) dx$ and an induced notion of norm or length given by $\|f\| = \sqrt{\langle f, f \rangle}$.

A key observation is that the system of Haar wavelets $\{h_n, n \geq 0\}$ forms an orthonormal system; in other words, each h_n has unit norm, $\|h_n\| = 1$, and the system of functions is pairwise orthogonal, $\langle h_m, h_n \rangle = 0$ if $m \neq n$. That these functions are properly normalised is clear as h_0 and h_1 trivially have unit norm, $\|h_0\| = \|h_1\| = 1$, and, for $n > 1$, a change of variable inside the integral reduces consideration to the mother wavelet h_1 . In detail,

$$\|h_{2^j+k}\|^2 = \int_0^1 h_{2^j+k}(x)^2 dx = \int_0^1 2^j h_1(2^j(x-k2^{-j}))^2 dx = \int_{\max\{0, -k\}}^{\min(1, 2^j-k)} h_1(y)^2 dy = 1,$$

the limits of integration in the penultimate step reducing to the unit interval for every choice of j and k . Orthogonality follows by inspection. (Or, if the reader must, by induction via the observation that the k th dyadic interval of width 2^{-j} in row j is partitioned into two dyadic intervals of width 2^{-j-1} by the $2k$ th and $(2k+1)$ th dyadic intervals in row $j+1$. The support of every wavelet in row $j+1$ is hence in an interval of constancy of width 2^{-j-1} of each wavelet in row j .)

The Haar system may hence be thought of as establishing a system of orthogonal “coordinate directions” in the space $L^2[0, 1]$. For any square-integrable f , it is now natural to interpret the inner product $\hat{f}_n = \langle f, h_n \rangle$ as the projection of f in the n th coor-

dinate direction, the partial sum $S_N(f, x) = \sum_{n=0}^N \hat{f}_n h_n(x)$ representing the projection of f onto the subspace spanned by the Haar wavelets h_0, h_1, \dots, h_N . The geometric intuition that these projections capture the essence of f in the Haar system is supported by Parseval's equation. Suppose there exists a sequence of real numbers $\{\lambda_n, n \geq 0\}$ such that, as $N \rightarrow \infty$, the partial sum $f_N(x) = \sum_{n=0}^N \lambda_n h_n(x)$ in the linear span of the Haar functions h_0, h_1, \dots, h_N converges in mean-square to f , that is to say, $\|f - f_N\| \rightarrow 0$ as $N \rightarrow \infty$. Then $\|f - S_N(f, \cdot)\| \rightarrow 0$ as well and, moreover, $\|f\|^2 = \sum_{n=0}^{\infty} \hat{f}_n^2$. The final identity is Parseval's equation which says that if $\{f_N\}$ converges to f in mean-square then the energy of f is preserved in its projections onto the Haar coordinate system. The reader experienced in Fourier analysis will know that showing convergence results of this stripe is in general a ticklish business requiring delicate arguments. The extreme localisation of the Haar system, however, makes these arguments almost trite.

For $j \geq 0$ and $0 \leq k \leq 2^j - 1$, let $\chi_{jk}(x)$ denote the indicator for the dyadic interval $[k2^{-j}, (k+1)2^{-j})$, that is $\chi_{jk}(x) = 1$ if $k2^{-j} \leq x < (k+1)2^{-j}$ and is zero otherwise. An easy induction shows that these dyadic indicators may be written as a linear combination of Haar functions. Indeed, for $j = 0$ we have the trivial $\chi_{00} = h_0$, while for $j = 1$ we have the almost as obvious $\chi_{10} = \frac{1}{2}(h_0 + h_1)$ and $\chi_{11} = \frac{1}{2}(h_0 - h_1)$. The induction is completed by the observation that $\chi_{j+1,2k} = \frac{1}{2}(\chi_{jk} + 2^{-j/2}h_{2^{j+k}})$ and $\chi_{j+1,2k+1} = \frac{1}{2}(\chi_{jk} - 2^{-j/2}h_{2^{j+k}})$.

Now any uniformly continuous function f on the unit interval may be uniformly well approximated by a sequence of step functions

$$f_j(x) = f(0)\chi_{j0}(x) + f(2^{-j})\chi_{j1}(x) + \dots + f(k2^{-j})\chi_{jk}(x) + \dots + f(1 - 2^{-j})\chi_{j,2^j-1}(x),$$

obtained by partitioning the unit interval into 2^j dyadic subintervals. In consequence, we have $|f(x) - f_j(x)| < \epsilon$ and, accordingly, $\|f - f_j\|^2 = \int_0^1 (f(x) - f_j(x))^2 dx < \epsilon^2$, for any $\epsilon > 0$ and all sufficiently large j . And so $\{f_j, j \geq 0\}$ converges to f in mean-square. As each f_j may be represented as a linear combination of Haar functions, it follows by the mean-square approximation theorem that *if f is uniformly continuous then the sequence $S_N(f, \cdot) = \sum_{n=0}^N \langle f, h_n \rangle h_n$ converges to f in mean-square*. The simple nature of the demonstration will be appreciated by the reader experienced in Fourier analysis who has seen the finesse required to show that the Fourier basis is complete.

The fact that the dyadic indicators lie in the linear span of the Haar system fuels an amusing consequence of Parseval's equation. Suppose $[0, s)$ and $[0, t)$ are subintervals of the unit interval with dyadic endpoints, that is, s and t are integer multiples of 2^{-j} for some j . Write f and g for the indicators of $[0, s)$ and $[0, t)$, respectively. That is to say, $f(x) = 1$ for $0 \leq x < s$ and zero otherwise, and $g(x) = 1$ for $0 \leq x < t$ and zero otherwise. The indicators of these intervals may be written as a sum of dyadic indicators and hence as a linear combination of Haar functions. By Parseval's equation it follows immediately that $\langle f, g \rangle = \sum_{n=0}^{\infty} \langle f, h_n \rangle \langle g, h_n \rangle$. Now, on the one hand we have $\langle f, g \rangle = \min\{s, t\}$, while on the other $\langle f, h_n \rangle = \int_0^s h_n(x) dx$ and $\langle g, h_n \rangle = \int_0^t h_n(x) dx$. A curious identity peeks shyly out:

$$\min\{s, t\} = \sum_{n=0}^{\infty} \int_0^s h_n(x) dx \int_0^t h_n(x) dx. \quad (10.1)$$

The identity may be extended to any $s, t \in [0, 1]$ as we may select intervals with dyadic endpoints as close as desired to s and t . But we digress.

11 A bare hands construction

Suppose $\{Z_n, n \geq 0\}$ is a sequence of independent, normal random variables each with mean zero and variance one. (The careful reader who worries about whether such a sequence can be constructed will find reassurance in Section XII.9.) We attempt to construct a sample function of a Brownian motion process on the unit interval $[0, 1]$ via a series of the form

$$X(t) = \sum_{n=0}^{\infty} Z_n \vartheta_n(t) \quad (11.1)$$

for some suitably chosen sequence of functions $\{\vartheta_n, n \geq 0\}$. It is plausible that as we are dealing with an infinite sum of normals that the result is also normal. But if this is the case then the distribution is entirely specified by the covariances. Proceeding boldly to exchange expectation and summation we obtain

$$\text{Cov}(X(s), X(t)) \stackrel{?}{=} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} E(Z_m Z_n) \vartheta_m(s) \vartheta_n(t) = \sum_{n=0}^{\infty} \vartheta_n(s) \vartheta_n(t)$$

as $E(Z_m Z_n) = 0$ if $m \neq n$ and $E(Z_n^2) = \text{Var } Z_n = 1$. Of course, the step marked “?” requires justification but, ignoring the niceties of convergence for the nonce, let us see what we can conclude from the purely formal manipulation. If we are to obtain a representation of standard Brownian motion the left-hand side must equal $\min\{s, t\}$. The apparently idle identity (10.1) now suggests a speculative choice.

Beginning with the orthogonal Haar system $\{h_n, n \geq 0\}$, we build a corresponding sequence of functions $\{\vartheta_n, n \geq 0\}$ with support in the unit interval by setting $\vartheta_n(t) = \int_0^t h_n(x) dx$ for each $n \geq 0$. These functions are readily determined. If $n = 0$ then $\vartheta_0(t) = t$ for $0 \leq t < 1$. Likewise, for $n = 1$, we obtain $\vartheta_1(t) = t$ for $0 \leq t < 1/2$ and $\vartheta_1(t) = 1 - t$ for $1/2 \leq t < 1$. For the remaining cases, a simple change of variable inside the integral shows that $\vartheta_{2^j+k}(t) = 2^{-j/2} \vartheta_1(2^j(t - k2^{-j}))$ for each $j \geq 0$ and $0 \leq k \leq 2^j - 1$. Most of what is salient is captured in Figure 9: the sequence of functions $\{\vartheta_n, n \geq 0\}$ constitutes a new, triangular wavelet system with the graph of each ϑ_n forming a triangle with support in the same dyadic subinterval where h_n has support.

We have now identified a system of functions $\{\vartheta_n, n \geq 0\}$ as putative building blocks for a formal construction of Brownian motion on the unit interval. Of course, many things will have to be verified before we can conclude with any confidence that we have been successful. The first item of business is to ask whether the series (11.1) converges at all. It is key that the normal tails die very quickly.

LEMMA *For every $0 < \delta < 1$, we may select $\zeta = \zeta(\delta)$ so that the probability that none of the events $|Z_n| \geq \sqrt{2\zeta \log n}$ occur (where n ranges over all integers ≥ 2) is at least $1 - \delta$.*

PROOF: The normal tail bound of Lemma VI.1.3 shows that $P\{|Z_n| \geq \sqrt{2\zeta \log n}\} \leq e^{-2\zeta \log(n)/2} = n^{-\zeta}$. By subadditivity, it follows that

$$P\left(\bigcup_{n=2}^{\infty} \{|Z_n| \geq \sqrt{2\zeta \log n}\}\right) \leq \sum_{n=2}^{\infty} P\{|Z_n| \geq \sqrt{2\zeta \log n}\} \leq \sum_{n=2}^{\infty} n^{-\zeta}.$$

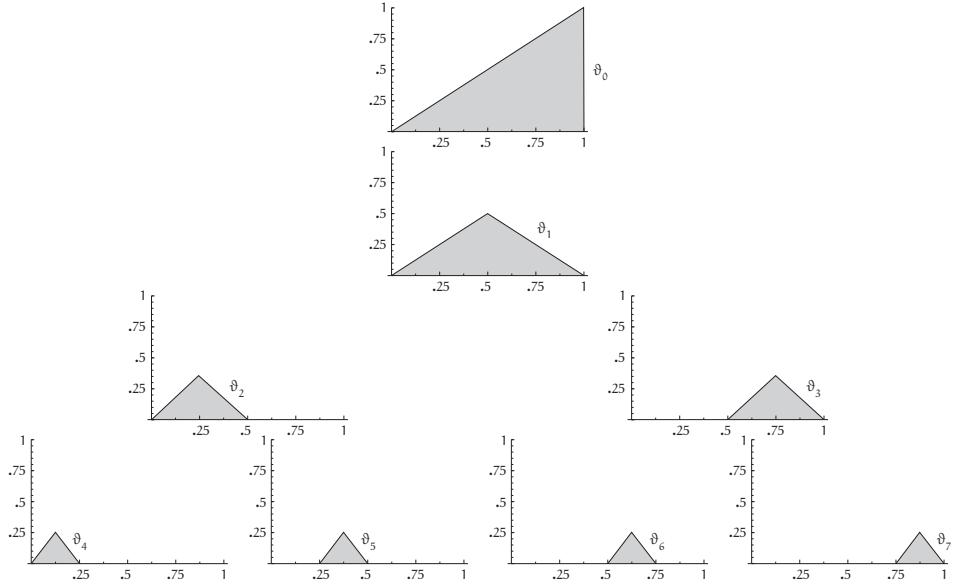


Figure 9: The triangular wavelet system inherited from the Haar wavelets.

By grouping the terms of the sum on the right in blocks of powers of two, we may express it in the form $(2^{-\zeta} + 3^{-\zeta}) + (4^{-\zeta} + 5^{-\zeta} + 6^{-\zeta} + 7^{-\zeta}) + (8^{-\zeta} + \dots + 15^{-\zeta}) + \dots$, and, as the leading term is dominant in each group, we see that

$$\sum_{n=2}^{\infty} n^{-\zeta} \leq 2^{-\zeta} \cdot 2 + 4^{-\zeta} \cdot 4 + 8^{-\zeta} \cdot 8 + \dots = \sum_{k=1}^{\infty} 2^{(-\zeta+1)k} = \frac{2^{-\zeta+1}}{1 - 2^{-\zeta+1}}.$$

If $\zeta \geq 2$ the upper bound on the right is no larger than $2^{-\zeta+2}$ and, in particular, if we select $\zeta = \zeta(\delta) = -\log_2(\delta/4)$ then the right-hand side does not exceed δ . ▶

In order to examine the convergence of (11.1), the recursive wavelet construction suggests that we begin by rewriting the series in the form

$$X(t) = Z_0 \vartheta_0(t) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} Z_{2j+k} \vartheta_{2j+k}(t). \quad (11.1')$$

The functions $\vartheta_n(t)$ are continuous and decay rapidly with j . Indeed, $0 \leq \vartheta_0(t) \leq 1$ and, as $0 \leq \vartheta_1(t) \leq 1/2$, we have $0 \leq \vartheta_{2j+k}(t) \leq 2^{-1-j/2}$ for $j \geq 0$ and $0 \leq k \leq 2^j - 1$. The key fact that now enables us to verify that (11.1') converges prettily is the extreme localisation of the triangular wavelets: *for any $j \geq 0$, any given value t lies in the support of precisely one of the 2^j functions $\vartheta_{2j}, \vartheta_{2j+1}, \dots, \vartheta_{2j+2^j-1}$.* To put this potent observation into action, fix any $\delta > 0$ and any positive integer J and set $N = 2^J$. In view of the preceding lemma, it follows that, for any $0 < \delta < 1$, we may select $\zeta = \zeta(\delta)$ so that, with probability at least $1 - \delta$, the tail of the series (11.1') is bounded by

$$\begin{aligned} \sum_{n=N}^{\infty} |Z_n| \vartheta_n(t) &= \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} |Z_{2^j+k}| \vartheta_{2^j+k}(t) \leq \sqrt{2\zeta} \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} \sqrt{\log(2^j + k)} \vartheta_{2^j+k}(t) \\ &\leq \sqrt{2\zeta \log 2} \sum_{j=J}^{\infty} \sqrt{j+1} \sum_{k=0}^{2^j-1} \vartheta_{2^j+k}(t) \leq \sqrt{2\zeta \log 2} \sum_{j=J}^{\infty} \sqrt{j+1} 2^{-1-j/2} \end{aligned}$$

as, for any $j \geq 0$, there is a non-zero contribution from at most one of the terms of the inner sum. The reader should note that the dependence on t has vanished in the upper bound. In view of the rapid extinction of the exponential in the summand, for any specified $\epsilon > 0$ and $\delta > 0$, we may hence select $J = J(\epsilon, \delta)$ uniformly in t so that the upper bound on the right is no larger than ϵ . (For instance, the observation $j+1 \leq 2^{j/2}$ for $j \geq 6$ allows us to bound the right-hand side by a geometric series.) As ϵ may be chosen arbitrarily small, it follows that, with probability at least $1 - \delta$, the sequence $X_N(t) = \sum_{n=0}^{N-1} Z_n \vartheta_n(t)$ converges uniformly to a limit. The partial sums $X_N(t)$ constitute a sequence of continuous functions for each realisation of the random variables $\{Z_n, n \geq 0\}$. As the uniform limit of a sequence of continuous functions is continuous,¹⁶ it follows that the sequence $X_N(t)$ converges to a limiting continuous function $X(t)$ with probability at least $1 - \delta$. As δ may be chosen arbitrarily small, it follows that *the series (11.1') does indeed converge to a continuous function with probability one.*

As (11.1') represents an “infinite” sum of normal variables it is tempting to declare “mission accomplished” and assert that $X(t)$ is a Gaussian process. We should temper our enthusiasm with some caution, however, as we have a limiting process at hand. The (high probability) uniform convergence of the series points the way.

Fix any $0 < t < 1$. For each $N \geq 1$, the partial sum $X_N(t) = \sum_{n=0}^{N-1} Z_n \vartheta_n(t)$ is a finite sum of normal variables, hence has impeccably normal credentials. Its mean is clearly zero and its variance is given by

$$\text{Var } X_N(t) = \mathbb{E}(X_N(t)^2) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \mathbb{E}(Z_m Z_n) \vartheta_m(t) \vartheta_n(t) = \sum_{n=0}^{N-1} \vartheta_n(t)^2.$$

The expression on the right is a partial sum of the convergent series $\sum_{n=0}^{\infty} \vartheta_n(t)^2$ which, by (10.1) applied to the case $s = t$, converges to t . It follows that $\text{Var } X_N(t) = t - \delta_N$ where $\delta_N = \delta_N(t) = \sum_{n=N}^{\infty} \vartheta_n(t)^2 \rightarrow 0$ as $N \rightarrow \infty$. For each t then, the partial sum $X_N(t)$ is normal with mean zero and variance $t - \delta_N \rightarrow t$. As the normal density varies continuously with respect to its variance viewed as a parameter, it is clear that

$$\frac{1}{\sqrt{t - \delta_N}} \phi\left(\frac{x}{\sqrt{t - \delta_N}}\right) \rightarrow \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) \quad (N \rightarrow \infty)$$

uniformly for all x in any closed and bounded interval. It follows that

$$\mathbb{P}\{X_N(t) \in \mathbb{I}\} = \int_{\mathbb{I}} \frac{1}{\sqrt{t - \delta_N}} \phi\left(\frac{x}{\sqrt{t - \delta_N}}\right) dx \rightarrow \int_{\mathbb{I}} \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx$$

as $N \rightarrow \infty$ for any closed and bounded interval \mathbb{I} .

¹⁶The reader who does not remember this fact will find it in Theorem XXI.2.3 in the Appendix.

Set $X'_N(t) = X(t) - X_N(t) = \sum_{n \geq N} Z_n \vartheta_n(t)$. Then for any $\epsilon > 0$ and $\delta > 0$ we may choose $N = N(\epsilon, \delta)$ so large that $|X'_N(t)| < \epsilon$ with probability $\geq 1 - \delta$. Now, by conditioning on $X'_N(t)$, we have

$$\begin{aligned} P\{a < X(t) < b\} &= P\{a - X'_N(t) < X_N(t) < b - X'_N(t)\} \\ &= P\{a - X'_N(t) < X_N(t) < b - X'_N(t) \mid |X'_N(t)| < \epsilon\} P\{|X'_N(t)| < \epsilon\} \\ &\quad + P\{a - X'_N(t) < X_N(t) < b - X'_N(t) \mid |X'_N(t)| \geq \epsilon\} P\{|X'_N(t)| \geq \epsilon\}. \end{aligned}$$

As $X_N(t)$ and $X'_N(t)$ are determined by disjoint sets of independent variables, hence are independent, the expression on the right is bounded below by $P\{a + \epsilon < X_N(t) < b - \epsilon\} \cdot (1 - \delta) \geq P\{a + \epsilon < X_N(t) < b - \epsilon\} - \delta$, while it is bounded above by $P\{a - \epsilon < X_N(t) < b + \epsilon\} \cdot 1 + \delta$. It follows that

$$P\{a + \epsilon < X_N(t) < b - \epsilon\} - \delta \leq P\{a < X(t) < b\} \leq P\{a - \epsilon < X_N(t) < b + \epsilon\} + \delta. \quad (11.2)$$

Proceeding to the limit as $N \rightarrow \infty$ yields

$$\int_{a+\epsilon}^{b-\epsilon} \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx - \delta \leq P\{a < X(t) < b\} \leq \int_{a-\epsilon}^{b+\epsilon} \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx + \delta,$$

and the reader may sense a Lévy-like sandwich in the offing. The integrals on the left and on the right may be brought as close to each other as desired and, in particular, for a sufficiently small choice of $\epsilon = \epsilon(\delta)$, each of the bookend integrals will differ from $\int_a^b \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx$ in no more than δ . It follows that for every choice of $0 < \delta < 1$, we have

$$\left| P\{a < X(t) < b\} - \int_a^b \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx \right| \leq 2\delta,$$

and as δ may be chosen arbitrarily small we are led to the inexorable conclusion that

$$P\{a < X(t) < b\} = \int_a^b \frac{1}{\sqrt{t}} \phi\left(\frac{x}{\sqrt{t}}\right) dx$$

for any $a < b$. Thus, for every $0 < t < 1$, $X(t)$ is normal with mean zero and variance t .

To show that all the finite-dimensional distributions of the process (11.1') are normal requires little more than to cast the argument into vector language. Letting t_1, \dots, t_k be any fixed points in the unit interval $(0, 1)$ we now consider the natural vector analogues of our earlier notation, $\mathbf{X} = (X(t_1), \dots, X(t_k))$, $\mathbf{X}_N = (X_N(t_1), \dots, X_N(t_k))$, and $\mathbf{X}'_N = \mathbf{X} - \mathbf{X}_N$.

For each i and j , the covariance of the random variables $X_N(t_i)$ and $X_N(t_j)$ is given by a process similar to the computation of the variance of $X_N(t)$ to be

$$\gamma_N(i, j) = \text{Cov}(X_N(t_i), X_N(t_j)) = E(X_N(t_i)X_N(t_j)) = \sum_{n=0}^{N-1} \vartheta_n(t_i)\vartheta_n(t_j).$$

Writing $\gamma(i, j) = \min\{t_i, t_j\}$, in view of (10.1), $\gamma_N(i, j) \rightarrow \gamma(i, j)$ for each i and j . It follows that the matrix of covariances $\text{Cov } \mathbf{X}_N = \Gamma_N = [\gamma_N(i, j)]$ converges componentwise to the matrix $\Gamma = [\gamma(i, j)]$.

Now \mathbf{X}_N is obtained by a linear transformation of normal variables, hence is normal, centred at the origin, and with matrix of covariances $\Gamma_N = [\gamma_N(i, j)]$. Introduce the nonce notation $f_{\Gamma_N}(x)$ and $f_{\Gamma}(x)$ for k -variate normal densities, each centred at the origin, and with covariance matrices Γ_N and Γ , respectively. As the multivariate normal varies smoothly with the covariance parameters, it follows that $f_{\Gamma_N}(x) \rightarrow f_{\Gamma}(x)$ uniformly over any closed and bounded region of \mathbb{R}^k and hence

$$\mathbf{P}\{\mathbf{X}_N \in \mathbb{I}\} \rightarrow \int_{\mathbb{I}} f_{\Gamma}(x) dx = \int_{\mathbb{I}} \frac{1}{(2\pi)^{k/2} \det(\Gamma)^{1/2}} \exp(-\frac{1}{2}x\Gamma^{-1}x^T) dx \quad (N \rightarrow \infty)$$

for any closed and bounded region \mathbb{I} in \mathbb{R}^k .

Suppose $\mathbf{a} = (a_1, \dots, a_k)$ and $\mathbf{b} = (b_1, \dots, b_k)$ are vectors in \mathbb{R}^k . As a notational convention we interpret the vector inequality $\mathbf{a} < \mathbf{b}$ to mean that the inequality holds componentwise.

To return to the analysis, we may select N sufficiently large so that each of the components of \mathbf{X}'_N is bounded absolutely by ϵ with probability at least $1 - \delta$. Proceeding by conditioning on \mathbf{X}'_N as before, for every choice of $\mathbf{a} < \mathbf{b}$, we obtain

$$\mathbf{P}\{\mathbf{a} + \epsilon \mathbf{1} < \mathbf{X}_N < \mathbf{b} - \epsilon \mathbf{1}\} - \delta \leq \mathbf{P}\{\mathbf{a} < \mathbf{X} < \mathbf{b}\} \leq \mathbf{P}\{\mathbf{a} - \epsilon \mathbf{1} < \mathbf{X}_N < \mathbf{b} + \epsilon \mathbf{1}\} + \delta \quad (11.2')$$

which is the k -dimensional version of (11.2). Here $\mathbf{1}$ stands for the k -dimensional vector for all of whose components are 1. The steps now mirror those in one dimension—all that is needed is to interpret variables as vectors, inequalities as vector inequalities, and integrals over \mathbb{R}^k . By passing to the limit as $N \rightarrow \infty$, (11.2') becomes

$$\int_{\mathbf{a} + \epsilon \mathbf{1}}^{\mathbf{b} - \epsilon \mathbf{1}} f_{\Gamma}(x) dx - \delta \leq \mathbf{P}\{\mathbf{a} < \mathbf{X} < \mathbf{b}\} \leq \int_{\mathbf{a} - \epsilon \mathbf{1}}^{\mathbf{b} + \epsilon \mathbf{1}} f_{\Gamma}(x) dx + \delta.$$

The integrals on the left and right may again be brought as close to $\int_{\mathbf{a}}^{\mathbf{b}} f_{\Gamma}(x) dx$ as desired and it follows that $|\mathbf{P}\{\mathbf{a} < \mathbf{X} < \mathbf{b}\} - \int_{\mathbf{a}}^{\mathbf{b}} f_{\Gamma}(x) dx| \leq 2\delta$ for a sufficiently small choice of $\epsilon = \epsilon(\delta)$. As δ may be chosen arbitrarily small, it follows that $\mathbf{X} = (X(t_1), \dots, X(t_k))$ is normal with mean zero and covariance matrix Γ whose components are given by $\gamma(i, j) = \min\{t_i, t_j\}$. As Algernon Swinburne wrote in 1866 in his poem *The Garden of Proserpine*, even the weariest river winds somewhere safe to sea.

THEOREM 1 Suppose $\{Z_n, n \geq 0\}$ is a sequence of independent, normal random variables, each of mean 0 and variance 1 and let $\{h_n, n \geq 0\}$ denote the Haar wavelet system. Then the series

$$X(t) = \sum_{n=0}^{\infty} Z_n \int_0^t h_n(x) dx$$

represents a standard Brownian motion on the unit interval $[0, 1]$.

This elegant construction of Brownian motion is from Y. Meyer's beautiful account of wavelets in which I first became aware of the possibilities of the Haar system in this context.¹⁷

¹⁷Y. Meyer (translated and revised by R. D. Ryan), *Wavelets: Algorithms and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1993.

In some ways, the restriction of a Brownian motion path to the unit interval is a little unnatural—its natural habitat is the positive half of the real line with the properties extending without change. But it is now a simple matter to extend the construction to the half-line $[0, \infty)$ by simply stitching together independent copies of Brownian motion on unit intervals, each new segment beginning where the previous one left off.

THEOREM 2 *Let $X_0(t), X_1(t), X_2(t), \dots$ be a sequence of independent versions of standard Brownian motion on the unit interval $[0, 1]$. Then the process $X(t)$ defined on $[0, \infty)$ by*

$$X(t) = \begin{cases} X_0(t) & \text{if } 0 \leq t < 1, \\ X_0(1) + X_1(t - 1) & \text{if } 1 \leq t < 2, \\ X_0(1) + X_1(2 - 1) + X_2(t - 2) & \text{if } 2 \leq t < 3, \\ \dots & \dots \\ X_0(1) + X_1(2 - 1) + \dots + X_{n-1}(n - (n - 1)) + X_n(t - n) & \text{if } n \leq t < n + 1, \\ \dots & \dots \end{cases} \quad (11.3)$$

represents a standard Brownian motion process on the half-line \mathbb{R}^+ .

While the mechanism (11.3) can hardly make the concept clearer it can help in running a routine verification which I will leave to the reader.



12 The paths of Brownian motion are very kinky

Suppose $X(t)$ represents a standard Brownian motion on, say, the unit interval $[0, 1]$. Baked into the structure is the idea that the process has independent increments. Specifically, suppose $0 = t_0 < t_1 < \dots < t_n = 1$ and consider the increments $\Delta_1 = X(t_1) - X(t_0), \dots, \Delta_n = X(t_n) - X(t_{n-1})$. It is clear that $\Delta = (\Delta_1, \dots, \Delta_n)$ is obtained by a linear transformation of the normal vector $X = (X(t_1), \dots, X(t_n))$; indeed, $\Delta = XW$ where $W = [w_{jk}]$ is a matrix with non-zero components only along its diagonal and principal upper diagonal, $w_{kk} = 1$, $w_{k,k+1} = -1$, and $w_{jk} = 0$ if $j \neq k$. It follows that Δ is normal and centred at the origin, its distribution hence completely specified by its covariance matrix $\text{Cov}(\Delta) = W^T \text{Cov}(X)W$ which, written out explicitly,

$$\begin{pmatrix} + & 0 & 0 & 0 & \cdots & 0 \\ - & + & 0 & 0 & \cdots & 0 \\ 0 & - & + & 0 & \cdots & 0 \\ 0 & 0 & - & + & \cdots & 0 \\ \dots & 0 & 0 & 0 & \cdots & + \end{pmatrix} \begin{pmatrix} t_1 & t_1 & t_1 & t_1 & \cdots & t_1 \\ t_1 & t_2 & t_2 & t_2 & \cdots & t_2 \\ t_1 & t_2 & t_3 & t_3 & \cdots & t_3 \\ t_1 & t_2 & t_3 & t_4 & \cdots & t_4 \\ t_1 & t_2 & t_3 & t_4 & \cdots & t_n \end{pmatrix} \begin{pmatrix} + & - & 0 & 0 & \cdots & 0 \\ 0 & + & - & 0 & \cdots & 0 \\ 0 & 0 & + & - & \cdots & 0 \\ 0 & 0 & 0 & + & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & + \end{pmatrix},$$

is in the form of a product of matrices of very simple structure. Here $+$ and $-$ represent $+1$ and -1 , respectively. An easy evaluation now shows that the matrix product collapses into the matrix $\text{diag}(t_1 - t_0, t_2 - t_1, \dots, t_n - t_{n-1})$ so that $\text{Cov}(\Delta)$ is diagonal. It follows, as asserted following the definition in Section 9, that *the increments $\Delta_1, \dots, \Delta_n$ of the Brownian motion process are independent random variables with $\Delta_k \sim \mathcal{N}(0, t_k - t_{k-1})$ for $1 \leq k \leq n$.*

Consider now equispaced intervals with endpoints $t_k = k/n$ for $0 \leq k \leq n$ where we think of n as a large parameter. We write $\Delta_{nk} = X(k/n) - X((k-1)/n)$ for the corresponding increments, the subscript n introduced to keep the rôle of the asymptotic parameter firmly in view. The increments $\Delta_{n1}, \dots, \Delta_{nn}$ are independent with the common normal distribution $\mathcal{N}(0, 1/n)$. For large n , each increment has a standard deviation of $n^{-1/2}$ which is large compared to the interval width n^{-1} and so each increment is likely to jitter by an amount large compared to the time scale under consideration. The reader may be reminded of Perrin's observation, "The apparent mean speed of a grain during a given time varies in *the wildest way* in magnitude and direction, and does not tend to a limit as the time taken for observation decreases." We are led to the shrewd suspicion that the sample paths of Brownian motion must have a very irregular character. And indeed they do.

THEOREM *Excepting only a negligible collection of sample paths contained in a set of zero probability, the sample paths of Brownian motion are continuous everywhere but differentiable nowhere.*

It will be useful to build up some intuition before launching into the proof. If $X(\cdot)$ has a derivative at t then it cannot change very much in a vicinity of t . This leads us to consider increments in a small neighbourhood of t . To give ourselves a little room to manoeuvre, fix $\nu \geq 1$, and, for each k , consider the block of contiguous increments $\Delta_{n,k+1}, \dots, \Delta_{n,k+\nu}$. We will select ν appropriately once it becomes apparent how much flexibility is needed. The increments $\Delta_{n,k+1}, \dots, \Delta_{n,k+\nu}$ carry information about the process at points t in the immediate vicinity of k/n and we are led to consider the random variable $S_{nk} = \max\{|\Delta_{n,k+1}|, \dots, |\Delta_{n,k+\nu}|\}$ which represents the largest excursion of an increment in a vicinity of k/n . By allowing k to vary over all values from 0 to $n-\nu$ we cover all points in the unit interval and this will cater to the possibility that the target point t may lie anywhere in the interval. And so we finally come to a consideration of the random variable $T_n = \min\{S_{n0}, S_{n1}, \dots, S_{n,n-\nu}\}$ which represents the least of these large excursions. If $X(t)$ is differentiable at t then S_{nk} must be small for the value of k for which k/n is near t , whence T_n must also be small.

We may turn this rough analysis to account by a sample point-by-sample point consideration of the process. Each sample point ω of the chance experiment generating the Brownian motion process engenders a specific sample path $X(t) = X^\omega(t)$. [In the representation (11.1), we may take ω to range over the family of real-valued sequences $(z_n, n \geq 0)$ corresponding to realisations of the sequence $(Z_n, n \geq 0)$.] While a given sample path $X^\omega(\cdot)$ may, or may not, have a derivative at a given point t , the upper and lower derivatives from the right given, respectively, by

$$D_u^\omega(t) = \limsup_{h \downarrow 0} \frac{X(t+h) - X(t)}{h} \quad \text{and} \quad D_l^\omega(t) = \liminf_{h \downarrow 0} \frac{X(t+h) - X(t)}{h}$$

are well-defined, if possibly infinite. For each $M > 0$, let $\mathcal{E}^{(M)}$ denote the set of sample paths $X^\omega(\cdot)$ for which $|D_u^\omega(t)| < M$ and $|D_l^\omega(t)| < M$ for some $0 \leq t < 1$. As M increases, say, over the positive integers, $\mathcal{E}^{(M)}$ increases to a limit set $\mathcal{E} = \bigcup_M \mathcal{E}^{(M)}$ consisting of the collection of all sample paths $X^\omega(\cdot)$ for which $D_u^\omega(t)$ and $D_l^\omega(t)$ are both finite for some t . The theorem will be proved if we can show that the set \mathcal{E} is negligible, that is to say, contained in a set of sample paths of zero probability.

Fix any M and a sample path $X^\omega(\cdot)$ in $\mathcal{E}^{(M)}$. Select t such that $|D_u^\omega(t)| < M$ and $|D_l^\omega(t)| < M$. Then there exists $\delta > 0$ (determined by the choice of M , $X^\omega(\cdot)$, and t) so that $|X^\omega(t+h) - X^\omega(t)| \leq Mh$ for all $0 \leq h \leq \delta$. This is the key: the selected sample path can vary at most linearly in a vicinity of t .

With block size $\nu \geq 1$ fixed, select n so large that $(\nu+1)/n < \delta$ and pick for consideration the unique value k for which $(k-1)/n \leq t < k/n$. By choice of n , the block of ν intervals $[k/n, (k+1)/n], \dots, [(k+\nu-1)/n, (k+\nu)/n]$ is contained within the interval $[t, t+\delta]$ and, in particular, $0 < (k+m)/n - t \leq (\nu+1)/n < \delta$ for $1 \leq m \leq \nu$. The increment of the sample path $X^\omega(\cdot)$ over the interval $[(k+m-1)/n, (k+m)/n]$ may now be bounded via the triangle inequality by

$$|\Delta_{n,k+m}^\omega| = |X^\omega\left(\frac{k+m}{n}\right) - X^\omega\left(\frac{k+m-1}{n}\right)| \leq |X^\omega\left(\frac{k+m}{n}\right) - X^\omega(t)| + |X^\omega(t) - X^\omega\left(\frac{k+m-1}{n}\right)| \\ \leq \left(\frac{k+m}{n} - t\right)M + \left(\frac{k+m-1}{n} - t\right)M \leq \frac{2(\nu+1)M}{n}$$

for $1 \leq m \leq \nu$. It follows that $S_{n,k}^\omega = \max\{|\Delta_{n,k+1}^\omega|, \dots, |\Delta_{n,k+\nu}^\omega|\} \leq 2(\nu+1)M/n$, and hence also $T_n^\omega = \min\{S_{n,0}^\omega, S_{n,1}^\omega, \dots, S_{n,n-\nu}^\omega\} \leq 2(\nu+1)M/n$, as the uniform bound for the size of the increments in the block implies that the largest excursion within the block has the same bound. We have shown that if $X^\omega(\cdot) \in \mathcal{E}^{(M)}$, then $T_n^\omega \leq 2(\nu+1)M/n$ for all sufficiently large n . Matters are now reduced to a purely probabilistic question that we can set about answering.

The random variable T_n ranges across the values T_n^ω as ω sweeps through all sample points. Let $A_n^{(M)}$ denote the event $\{T_n \leq 2(\nu+1)M/n\}$. Then

$$\mathbf{P}(A_n^{(M)}) \leq (n-\nu) \mathbf{P}\{S_{n,0} \leq \frac{2(\nu+1)M}{n}\} = (n-\nu) \mathbf{P}\{|\Delta_{n,1}| \leq \frac{2(\nu+1)M}{n}\}^\nu$$

where, as $S_{n,0}, S_{n,1}, \dots, S_{n,n-1}$ have a common distribution, the first step follows by Boole's inequality, and the second step follows because the increments $\Delta_{n,1}, \dots, \Delta_{n,\nu}$ are independent with a common distribution. As the increment $\Delta_{n,1}$ is normal with mean zero and variance n^{-1} , we have

$$\mathbf{P}\left\{|\Delta_{n,1}| \leq \frac{2(\nu+1)M}{n}\right\} = \int_{-\infty}^{\infty} \frac{1}{n^{-1/2}} \phi\left(\frac{x}{n^{-1/2}}\right) \mathbf{1}\left(|x| < \frac{2(\nu+1)M}{n}\right) dx \\ \stackrel{(y=\sqrt{n}x)}{=} \int_{-\infty}^{\infty} \phi(y) \mathbf{1}\left(|y| < \frac{2(\nu+1)M}{n^{1/2}}\right) dy < \frac{4(\nu+1)M}{n^{1/2}}$$

as the standard normal density $\phi(\cdot)$ is bounded above by 1. We have hence shown that

$$\mathbf{P}(A_n^{(M)}) < (n-\nu) [4(\nu+1)Mn^{-1/2}]^\nu < C(\nu, M)n^{-\nu/2+1}$$

for an absolute positive constant $C(\nu, M)$ determined by ν and M . Any block size $\nu > 2$ gives a non-trivial bound but if we select a slightly larger value of ν the convergence to zero is very rapid. In particular, if we select $\nu = 5$ then $\mathbf{P}(A_n^{(M)}) < C(5, M)n^{-3/2}$ and the probability converges to zero so rapidly indeed that $\sum_n \mathbf{P}(A_n^{(M)})$ converges. By the first Borel-Cantelli lemma of Section IV.4, it follows that $\mathbf{P}(A_n^{(M)} \text{ i.o.}) = 0$.

If we write $A^{(M)}$ for the event that $A_n^{(M)}$ occurs infinitely often, that is, $A^{(M)} = \limsup_n A_n^{(M)} = \bigcap_m \bigcup_{n \geq m} A_n^{(M)}$, we see hence that the sample points engendering the sample paths in $\mathcal{E}^{(M)}$ are contained in the zero probability set $A^{(M)}$. Allowing M to

range over the positive integers, the union of the zero probability sets $A^{(M)}$ is another zero probability set A ; indeed, $0 \leq P(A) \leq \sum_M P(A^{(M)}) = 0$. As A contains all sample points ω engendering sample paths $X^\omega(\cdot)$ for which the upper and lower right-sided derivatives, $D_u^\omega(t)$ and $D_l^\omega(t)$, are both finite for some t , it is clear that any sample path for which there exists a derivative anywhere is engendered by a sample point in A . This concludes the proof of the theorem.

The verbal contortions in the preamble of the theorem are prompted by the realisation that, if B denotes the set of sample points ω for which $X^\omega(\cdot)$ has a derivative *somewhere*, then nothing in our provisions guarantees that B is measurable, that is to say, an event. But B is contained in a bona fide measurable set A of zero probability.

There are several subtle ideas in the proof that repay careful study. The idea of arguing sample point by sample point frequently clarifies the essence of the problem; the consideration of upper and lower derivatives finesse worries about whether the derivative actually exists at a point—a philosophically similar idea was seen in Section 7 in the consideration of second differences instead of second derivatives which were *a priori* guaranteed to exist—and the consideration of a block of independent increments drives down probabilities sufficiently quickly to overcome an increasing number of possibilities. A subtle use of the last idea is turned to great effect in Kolmogorov’s proof of the strong law of large numbers as the reader will see in Section XVI.9.

We could go on from here to prove a variety of interesting facts about Brownian motion but we shall not. The reader interested in pursuing the subject further will find a wealth of detail in Kahane’s excellent monograph.¹⁸

13 Problems

1. *Polar coordinates.* Suppose X and Y are independent, standard normal random variables. Evaluate $E(X^2/(X^2 + Y^2))$. [Hint: Example VII.9.3 will help.]

2. *A counterintuitive result for the normal density.* For any $-1 < \rho < 1$, consider the function $f(x_1, x_2) = \phi(x_1, x_2; \rho) + g(x_1, x_2)$ where g is a non-zero function with skew symmetry in alternating quadrants of the plane, $g(-x_1, x_2) = g(x_1, -x_2) = -g(x_1, x_2)$, and satisfying $|g(x_1, x_2)| \leq \phi(x_1, x_2; \rho)$. (For instance, for any choice of $0 < t \leq 1$, it will suffice to select $g(x_1, x_2) = t \operatorname{sgn}(x_1) \operatorname{sgn}(x_2) \phi(x_1, x_2; \rho)$ where $\operatorname{sgn}(x)$ is the *signum function* taking value $+1$ if $x > 0$, -1 if $x < 0$, and 0 if $x = 0$.) Then f is not normal but its marginals are.

3. *Asymptotic expansion.* Show that $\Phi(-x) = \sum_{k=0}^{n-1} (-1)^k c_k x^{-(2k+1)} \phi(x) + \xi_n(x)$ where, for every $n \geq 1$, $\xi_n(x)/(x^{-(2n-1)} \phi(x)) \rightarrow 0$ as $x \rightarrow \infty$. [Hint: Integrate by parts.]

4. *Rice’s density.* Suppose X_1 and X_2 are independent normal random variables with means μ_1 and μ_2 , respectively, and common variance σ^2 . Let $R = \sqrt{X_1^2 + X_2^2}$. Write $s = \sqrt{\mu_1^2 + \mu_2^2}$ and let I_0 be the zeroth-order modified Bessel function of the first kind given by the integral equation $I_0(z) = \frac{1}{\pi} \int_0^\pi e^{z \cos \theta} d\theta$. Show that R has *Rice’s density* with support in the positive half-line $r \geq 0$ where it takes the form $f(r) =$

¹⁸J-P. Kahane, *Some Random Series of Functions*, Second Edition. Cambridge: Cambridge University Press, 1993.

$\frac{r}{\sigma^2} e^{-(s^2+r^2)/2\sigma^2} I_0\left(\frac{rs}{\sigma^2}\right)$. The quantity s represents the non-centrality parameter of the density. What are the mean and variance of R ?

5. Suppose X_1, X_2, \dots is a sequence of independent, standard normal random variables. For each n , let $S_n = X_1 + \dots + X_n$. For $m < n$ determine the (joint) distribution of the random pair (S_m, S_n) . Determine the conditional density of S_m given that $S_n = t$.

6. In the preceding problem write $T_n = X_1^2 + \dots + X_n^2$. For $m < n$ determine the conditional density of T_m given that $T_n = t$.

7. Suppose X_1 and X_2 are independent, standard normal random variables. For each of the four cases illustrated in Figure 10, express the probability that $\mathbf{X} = (X_1, X_2)$ lies in the shaded region shown in terms of the normal d.f. Φ .

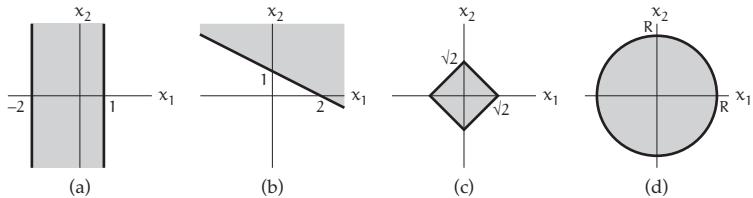


Figure 10: Regions of integration for Problem 7.

8. Suppose $\mathbf{X} = (X_1, X_2, X_3)$ has a normal density centred at the origin and let

$$A = \text{Cov}(\mathbf{X}) = \begin{pmatrix} 3 & 3 & 0 \\ 3 & 5 & 0 \\ 0 & 0 & 6 \end{pmatrix}.$$

(a) Determine the density of $Y = X_1 + 2X_2 - X_3$. (b) Determine the density of $\mathbf{Z} = (Z_1, Z_2, Z_3)$ where $Z_1 = 5X_1 - 3X_2 - X_3$, $Z_2 = -X_1 + 3X_2 - X_3$, and $Z_3 = X_1 + X_2$.

9. Suppose $\mathbf{X} = (X_1, X_2, X_3)$ has a normal density with covariance matrix

$$A = \text{Cov}(\mathbf{X}) = \begin{pmatrix} 1 & 0.9 & c \\ 0.9 & 1 & 0.9 \\ c & 0.9 & 1 \end{pmatrix}.$$

What is the smallest possible value that c can have?

10. Suppose X and Y are independent, standard normal. Let $Z = X + Y$. Determine the conditional density of Z given that $X > 0$ and $Y > 0$. Show thence that $E(Z | X > 0, Y > 0) = 2\sqrt{2}/\pi$.

11. Suppose (X, Y) is normal with zero means, unit variances, and $\text{Cov}(X, Y) = \rho$. Pass to polar coordinates $(X, Y) \mapsto (R, \Theta)$. Show that Θ has density $\sqrt{1-\rho^2}/(2\pi(1-2\rho\sin\vartheta\cos\vartheta))$ with support in $0 < \vartheta < 2\pi$ and hence that Θ is uniformly distributed if, and only if, $\rho = 0$. Conclude that $P\{XY > 0\} = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho$ and $P\{XY < 0\} = \frac{1}{\pi} \arccos \rho$.

12. For every $\nu > 0$, define the function $f_\nu(t) = \frac{\nu}{\sqrt{2\pi}t^{3/2}} \exp\left(-\frac{\nu^2}{2t}\right)$ with support in the positive half-line $t > 0$ only. Show that f_ν is the density of a positive random variable for each $\nu > 0$. [Hint: Suppose $X \sim N(0, \sigma^2)$. Consider X^{-2} .]

13. *Continuation, a one-sided stable distribution.* Show that $f_\mu * f_\nu(t) = f_{\mu+\nu}(t)$ for every $\mu > 0$ and $\nu > 0$. [Warning: A tedious integration is required.]

14. *Continuation.* Suppose $X \sim \mathcal{N}(0, \alpha^2)$ and $Y \sim \mathcal{N}(0, \beta^2)$ are independent normal variables. Show that $T = XY/\sqrt{X^2 + Y^2} \sim \mathcal{N}(0, \gamma^2)$ where $\gamma^{-1} = \alpha^{-1} + \beta^{-1}$. [Use the stability result of the previous problem.]

15. If Q is a symmetric matrix of order n with strictly positive eigenvalues show that the quadratic form xQx^T is strictly positive whenever x is not identically 0. The exponent of a normal density is hence strictly negative everywhere.

16. Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has a normal density in n dimensions. Show that there exists a unit vector $\mathbf{a} = (a_1, \dots, a_n)$ such that $\text{Var}(a_1 X_1 + \dots + a_n X_n) \geq \text{Var}(b_1 X_1 + \dots + b_n X_n)$ for all unit vectors $\mathbf{b} = (b_1, \dots, b_n)$. If $\mathbf{a} = (1, 0, \dots, 0)$ is such a vector then X_1 is independent of the rest of the X_j .

17. *Student's density.* Show that the variance of Student's density $s_\nu(t)$ decreases monotonically as ν increases.

18. *Continuation.* Show that Student's density $s_\nu(t)$ converges pointwise to the standard normal density $\phi(t)$ as $\nu \rightarrow \infty$.

19. Suppose X and Y are normal with mean zero, variance one, and covariance ρ . Let $Z = \max\{X, Y\}$. Show that $E(Z) = \sqrt{(1 - \rho)/\pi}$ and $E(Z^2) = 1$.

20. *A concentration result.* Suppose $\{X_j, j \geq 1\}$ is a sequence of independent, standard normal random variables. For each n , let $R_n = \sqrt{X_1^2 + \dots + X_n^2}$. Evaluate $E(R_n)$ and hence show that $E(R_n/\sqrt{n}) \rightarrow 1$ and $\text{Var}(R_n/\sqrt{n}) \rightarrow 0$. Conclude by the Chebyshev argument used in Sections V.6 and VIII.3 that $R_n/\sqrt{n} \xrightarrow{P} 1$ as $n \rightarrow \infty$. The limit here is in the sense of convergence in probability introduced in Section V.6 (see Section XII.5 for the full monty). For a strengthening, see Problem XVII.5.

21. *Diagonalisation.* Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has a normal density centred at the origin in \mathbb{R}^n with $A = \text{Cov}(\mathbf{X})$. Determine a linear transformation W on \mathbb{R}^n such that, with $\mathbf{Y} = \mathbf{X}W$, the components Y_1, \dots, Y_n are independent, standard normal random variables, each with mean zero and variance one.

22. *Continuation, simultaneous diagonalisation.* Suppose $\mathbf{X}' = (X'_1, \dots, X'_n)$ is a normal vector centred at the origin in \mathbb{R}^n with $A' = \text{Cov}(\mathbf{X}')$ and independent of \mathbf{X} . Determine a linear transformation \tilde{W} on \mathbb{R}^n such that, with $\mathbf{Z} = \mathbf{X}\tilde{W}$ and $\mathbf{Z}' = \mathbf{X}'\tilde{W}$, the components Z_1, \dots, Z_n of \mathbf{Z} are independent, standard normal random variables, each with mean zero and variance one, and the components Z'_1, \dots, Z'_n of \mathbf{Z}' are also independent normal random variables, not necessarily with the same variances.

23. *Hotelling's theorem.* Suppose $\mathbf{X}_r = (X_{r1}, \dots, X_{rn})$ ($1 \leq r \leq n$) are independent normal vectors in \mathbb{R}^n with zero means and common covariance matrix $A = [a_{jk}]$. Then $S_{jk} = \sum_{r=1}^n X_{rj} X_{rk} - \frac{1}{n} (\sum_{r=1}^n X_{rj}) (\sum_{r=1}^n X_{rk})$ and $T_{jk} = \sum_{r=1}^{n-1} X_{rj} X_{rk}$ have the same distribution. [Hint: Let $\mathbf{c}_1, \dots, \mathbf{c}_n$ be any system of orthonormal vectors in \mathbb{R}^n chosen so that \mathbf{c}_n is the vector all of whose components are $1/\sqrt{n}$. As r varies between 1 and n , form the vectors $\mathbf{Y}_r = (Y_{r1}, \dots, Y_{rn})$ with components specified by $Y_{rj} = \sum_{s=1}^n c_{rs} X_{sj}$. Argue that $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are independent normal vectors with the same distribution as $\mathbf{X}_1, \dots, \mathbf{X}_n$ and represent S_{jk} and T_{jk} in terms of the \mathbf{Y}_r .]

24. *A significance test in periodogram analysis.* A "trigonometric polynomial" with random coefficients of the form $Z(t) = \sum_{j=1}^n (X_j \cos \omega_j t + Y_j \sin \omega_j t)$ may be thought of

as an approximation of a stochastic process with continuous sample paths. Suppose $X_1, \dots, X_n, Y_1, \dots, Y_n$ are independent with a common normal distribution of mean zero and variance σ^2 . For each j , set $V_j = R_j^2 / (R_1^2 + \dots + R_n^2)$ where $R_j^2 = X_j^2 + Y_j^2$. Suppose $a > 0$. Determine the probability that all the V_j are less than a . An analysis of this type may be used to test whether a model fitting empirical observations has unearthed true periodic regularities in that the determined values V_1, \dots, V_n are atypical. [Hint: Use the result of Problem IX.21 to show that (V_1, \dots, V_n) is distributed as the spacings generated by $n - 1$ random points in the unit interval.]

25. A stationary Gaussian process. Suppose X and Y are independent, standard normal, and form the process $Z(t) = X \cos 2\pi t + Y \sin 2\pi t$. For every n , every choice of t_1, \dots, t_n , and every T , show that $(Z(t_1), \dots, Z(t_n))$ has the same distribution as $(Z(t_1 + T), \dots, Z(t_n + T))$. The finite-dimensional distributions of $Z(t)$ are hence normal and invariant with respect to choice of origin and we say in short that $Z(t)$ is a stationary Gaussian process. Determine the distribution of $(Z(0), Z(1/4), Z(1/2))$.

26. McCarthy's function. Let g be the plucked string function of period 4 defined over one period by $g(t) = 1 - |t|$ for $-2 \leq t \leq 2$. Let $f_n(t) = \sum_{k=1}^n 2^{-k} g(2^{2-k} t)$ for each $n \geq 1$. Show that the sequence f_n converges uniformly to a function f which is continuous everywhere and differentiable nowhere.¹⁹

27. The Brownian bridge. The triangular wavelet system $\{\vartheta_n(t), n \geq 0\}$ introduced in Section 11 satisfies $\vartheta_0(1-) = 1$ and $\vartheta_n(1-) = 0$ for $n \geq 1$. So, if we leave out the first term in the representation (11.1) for Brownian motion and define $U(t) = \sum_{n=1}^{\infty} Z_n \vartheta_n(t)$, we obtain a continuous process on the unit interval $[0, 1]$ satisfying $U(0) = U(1-) = 0$. This process is called a Brownian bridge and represents a process analogous to Brownian motion but which is tied down at its start and endpoints. (The imagery conjured up by the word bridge is certainly picturesque and now part of the settled terminology but it is not quite accurate.) Show that we can write $U(t) = X(t) - tX(1-)$ for $0 \leq t < 1$.

28. Continuation. Show that $\text{Cov}(U(s), U(t)) = s(1-t)$ for $0 \leq s \leq t < 1$.

29. Continuation. Suppose $X(t)$ is a standard Brownian motion on the unit interval $[0, 1]$ and let $V(t) = g(t)X(h(t))$. Find functions g and h so that $V(t)$ has the same covariance as a Brownian bridge.

30. Brownian motion via a Brownian bridge. Let $U(t)$ be the Brownian bridge process of Problem 27. Show that the process defined by $Y(t) = (1+t)U(\frac{t}{1+t})$ is a Brownian motion on the half-line $[0, \infty)$.

31. Verify that (11.3) has covariance $\text{Cov}(X(s), X(t)) = \min\{s, t\}$ and hence that $X(t)$ is indeed a standard Brownian motion on $[0, \infty)$.

32. Continuation, the scaled process. For any $a > 0$, show that the process $Y(t) = a^{-1/2}X(at)$ is a standard Brownian motion on $[0, \infty)$.

33. Continuation, the inverted process. Set $Z(0) = 0$ and $Z(t) = tX(1/t)$ for $t > 0$. Assuming that $Z(t)$ is continuous at $t = 0$, show that $Z(t)$ is a standard Brownian motion on $[0, \infty)$.

¹⁹This construction is due to J. McCarthy, "An everywhere continuous nowhere differentiable function", *American Mathematical Monthly*, vol. LX, no. 10, 1953.

Part B

FOUNDATIONS

Distribution Functions and Measure

Thus far we have explored what may be considered to be concepts in the elementary theory of probability. (Though, as the reader who has ploughed through the previous chapters will be aware, it would be a cardinal error to mistake the word elementary for any lack of subtlety or utility.) The various features that have been encountered hitherto, such as conditional probabilities, independent events, arithmetic distributions, and densities, are akin to recurring motifs that one encounters in an alien landscape; and the theorems we have developed are of necessity specialised to these cases. We now begin the work of considering what may be said from a more abstract point of view; and, in particular, to develop the foundations to the point that theorems of great generality and power can be naturally developed. To carry the metaphor of the mathematical tourist a little further, we now engage in attempting to decipher general characteristics and tendencies of the topography itself with a view to understanding the essential laws governing the mathematical landscape we find ourselves in.

c 1–3

§ 4, 5

We begin our exploration by investigating the fine structure of probability distributions on the line. The theory of measure plays an ineluctable rôle in this process. While there are several excellent books on measure they, of necessity, develop the subject in a context of much greater generality and focus on details that are not critical in our context. For our purposes it will be useful for us to focus on the measure-theoretic ideas central to probability and to lay them out clearly for reference. We develop this conceptual framework over this chapter and the next.

Measure-theoretic arguments are subtle, and consequently difficult, and the reader may well find that she agrees with T. W. Körner's assessment: "Mathematicians find it easier to understand and enjoy ideas which are clever rather than subtle. Measure theory is subtle rather than clever and so requires hard work to master." Fortunately, the part played by measure in most of the basic theory is to guarantee that limiting arguments carry through smoothly and it is certainly true that most of this book can be read without having to go through the niceties of the measure-theoretic proofs. (The theory validates intuitive

arguments and a vague “goes through generally” when faced with a limiting situation is all that will be required in most cases.) Accordingly, I have flagged the details of the measure-theoretic arguments—as well as other digressive material that has seemed fit to me to explore from time to time—in small print and in sections marked optional. If the reader wishes she can omit the flagged material on a first reading and return to sample it as necessity dictates and time or inclination allow.

Before launching into the chapter the reader should begin with a quick review of the elements of a probability space in Chapter I and, if she hasn’t already done so, read the dangerous bend Section I.8 which now becomes important. As preliminaries, little more is needed for background than some elementary facts about the real numbers that students are typically exposed to in an undergraduate calculus sequence. I have summarised the salient facts and notation in Section XXI.1 in the Appendix and the reader may wish to cast an eye over at least the main results before proceeding.

1 Distribution functions

The reader has built up some intuition for probabilities in real-line sample spaces through the first half of this book, and we use this as a jump-off point for the general theory. Probability measures on the real line are intimately connected with the theory of monotone functions.

DEFINITION A real-valued function $F(x)$ of a real variable x is called a *distribution function* (henceforth abbreviated *d.f.*) if it is right continuous, increasing, and has finite limits at $\pm\infty$ with $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.

We may take a harmless liberty with notation and write $F(-\infty) = 0$ and $F(\infty) = 1$ for the limiting values of the d.f. and we do so without further ado.

The importance of distribution functions stems from the fact that each d.f. induces a unique probability measure on the Borel sets of the real line and, conversely, every probability measure on the Borel sets can be associated with a d.f. We flesh out these assertions over the following sections.

Let $F: \mathbb{R} \rightarrow [0, 1]$ be a d.f. We begin by showing that F induces, in a certain natural sense, a unique probability measure F on the Borel sets of the line; that is to say, F induces a positive, countably additive set function F on the Borel sets so normalised as to map the real line \mathbb{R} to the value 1. For the purposes of this chapter we will use a bold font notation to indicate the set function induced by a d.f. to help visually separate the two kinds of functions—the d.f. $F(x)$ as an ordinary real-valued function of a real variable x , and the induced probability measure $F(\mathbb{I})$ on the Borel sets \mathbb{I} . This visual separation may be helpful initially to keep the concepts clearly in mind but it is not logically necessary and I will unburden the notation in the next chapter.

The intervals provide a natural starting point in the construction of a probability measure on the line. While from an analytical perspective it does not much matter which of the four basic interval types, open intervals (a, b) , closed intervals $[a, b]$, or half-closed (or half-open) intervals $[a, b)$ or $(a, b]$ we select for consideration, as a practical matter the half-closed intervals have the great advantage that the set difference of two half-closed intervals of a given type is another half-closed interval of that type, that is to say, the family of half-closed intervals of a given type is closed under set differences. (The reader will readily verify that the set differences of two open intervals or of two closed intervals are neither open nor closed.) While we may choose to focus on either type of half-closed interval—and indeed different authors select one or the other—popular usage now favours consideration of the intervals of the type $(a, b]$. There is no compelling reason to buck consensus in this regard and accordingly we write \mathcal{I} for the family of bounded intervals open at the left and closed at the right of the form $(a, b] = \{x : -\infty < a < x \leq b < +\infty\}$ and focus on intervals of this type. *Henceforth we shall reserve the term half-closed interval for a member of the family \mathcal{I} .*

Fix any $a \leq b$ and consider the half-closed interval $\mathbb{I} = (a, b]$. We begin by observing that F induces a set function on the half-closed intervals via

$$F(a, b] = F(b) - F(a).$$

As, by right continuity, $F(a, b] \rightarrow 0$ as $b \downarrow a$, it is natural to set $F(\emptyset) = 0$. Now $F(a, b] \geq 0$ as F is increasing. Furthermore, if $(a, b]$ is contained in $(a', b']$ then $a' \leq a \leq b \leq b'$, whence

$$F(a', b'] - F(a, b] = [F(b') - F(b)] + [F(a) - F(a')] \geq 0,$$

again by monotonicity of F . Accordingly, the set function F induced on the family of half-closed intervals \mathcal{I} is positive and monotone and already has the hallmarks of an incipient probability measure. That F can be extended to a *unique* probability measure on the Borel sets of the line is more than a little satisfying.

CARATHÉODORY'S EXTENSION THEOREM *Each d.f. F induces a unique probability measure \mathbf{F} on the Borel sets of the real line satisfying $\mathbf{F}(a, b] = F(b) - F(a)$ on each half-closed interval $(a, b]$.*

The simplest and most elegant approach to proof is via a device of Carathéodory. The proof is certainly lengthier and more technical than any encountered hitherto; and while no single element is particularly hard, the reader may well find that the corpus is a little daunting on first viewing. The technical details of the proof, however, are not needed elsewhere in the book and I have accordingly deferred its presentation to the last two sections of this chapter where the reader will see how to complete the construction of the measure \mathbf{F} and hence the proof of the theorem. The reader anxious to progress to the *res gestae*

may wish to skip on by after a quick glance at Sections 2 and 3 to the general definition of a random variable in the next chapter and reserve Carathéodory's construction for later study.

2 Measure and its completion

The only change we have to make to move from probability measure to a general theory of measure is to relax the normalisation requirement to permit possibly unbounded measures. While we will need little of the general theory of measure in this book it will be convenient to collect the definitions and principal properties in one place to permit the odd reference to unbounded measure without unnatural gyrations to force the settings into a probabilistic straitjacket.

To make provision for possibly unbounded measures, it is convenient to extend the real line to include the two symbols $+\infty$ and $-\infty$ with the natural conventions regarding sums and products involving $\pm\infty$: if x is any real number then $x \pm \infty = \pm\infty$, $x/\pm\infty = 0$, and $x \cdot (\pm\infty)$ is $\pm\infty$ if x is positive, 0 if x is zero, and $\mp\infty$ if x is negative; $(\pm\infty) + (\pm\infty) = \pm\infty$, $(\pm\infty)(\pm\infty) = +\infty$, and $(\pm\infty)(\mp\infty) = -\infty$. The real line equipped with the two additional symbols $+\infty$ and $-\infty$ is called the *extended real line* and denoted $\overline{\mathbb{R}}$.

We begin with an abstract space \mathcal{X} equipped with a σ -algebra of subsets \mathcal{M} . Any set A in \mathcal{M} is called a *measurable set*, the pair $(\mathcal{X}, \mathcal{M})$ called a *measurable space*, though the terminology is misleading as a measure is yet to be specified. Suppose μ is an extended real-valued set function on \mathcal{M} . In keeping with terminology for probability measure, we say that: μ is *positive* if $\mu(A) \geq 0$ for each $A \in \mathcal{M}$; μ is *additive* if $\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2)$ for all disjoint sets A_1 and A_2 in \mathcal{M} ; and μ is *continuous from above at \emptyset* if, whenever $\{A_n\}$ is a decreasing sequence of sets in \mathcal{M} , of which at least one satisfies $\mu(A_n) < \infty$, and for which $A_n \downarrow \bigcap_n A_n = \emptyset$, we have $\mu(A_n) \rightarrow \mu(\emptyset)$ as $n \rightarrow \infty$, or, in short, $\lim_n \mu(A_n) = \mu(\lim_n A_n) = \mu(\emptyset)$.

DEFINITION 1 A *measure* μ on \mathcal{M} is a set function that is extended real-valued, positive, additive, continuous from above at \emptyset , and satisfies $\mu(\emptyset) = 0$. We say that μ is *finite* if $\mu(\mathcal{X}) < \infty$ and that μ is *σ -finite* if there exists an increasing sequence of measurable sets \mathcal{X}_n , each of finite μ -measure, such that $\mathcal{X}_n \uparrow \mathcal{X}$.

Probability measures are obviously finite in view of the normalisation condition. The archetypal σ -finite measure is Lebesgue measure; the intervals $(-n, n)$ each have finite length and increase to the real line which has unbounded length.

THEOREM 1 Suppose μ is a measure on a σ -algebra \mathcal{M} . Then:

- 1) If A and B are elements of \mathcal{M} with $A \subseteq B$ then $\mu(A) \leq \mu(B)$, that is, μ is *monotone*.

- 2) If $\{A_n\}$ is an increasing sequence of sets in \mathcal{M} and $A = \bigcup_n A_n$ then $\lim_n \mu(A_n) = \mu(\lim_n A_n) = \mu(A)$, that is, μ is continuous from below at A .
- 3) If $\{A_n\}$ is a decreasing sequence of sets in \mathcal{M} of which at least one has finite measure and $A = \bigcap_n A_n$ then $\lim_n \mu(A_n) = \mu(\lim_n A_n) = \mu(A)$, that is, μ is continuous from above at A .
- 4) If $\{A_n\}$ is a countable family of disjoint sets in \mathcal{M} then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$, that is, μ is countably additive.
- 5) If $\{A_n\}$ is a countable family of sets in \mathcal{M} and $\bigcup_n A_n \supseteq A$ where A is also in \mathcal{M} then $\mu(A) \leq \sum_n \mu(A_n)$, that is, μ has the subadditive property.

The proofs of these assertions parallel those for probability measure and serve as a useful review exercise (though the reader will have to watch out for stray appeals to finiteness as a general measure is not necessarily finite).

The structure of measure is satisfactory but for an annoying gap: one would like to make the natural claim that all subsets of a set M of measure zero also have measure zero. But a given subset N of a measurable set $M \in \mathcal{M}$ is not necessarily itself measurable and hence may not have a measure attached to it by μ . This is easy to fix by enlarging the family of measurable sets slightly and extending the definition of μ to this family.

DEFINITION 2 The family $\overline{\mathcal{M}}$ (called the *completion* of \mathcal{M}) consists of all sets of the form $A \cup N$ where $A \in \mathcal{M}$ is any measurable set and N is a subset of any measurable set M with $\mu(M) = 0$.

It is clear that $\overline{\mathcal{M}}$ includes all the measurable sets because any measurable set $A \in \mathcal{M}$ may be written in the form $A \cup \emptyset$ and the empty set has measure zero.

THEOREM 2 *The completion $\overline{\mathcal{M}}$ is a σ -algebra containing all the measurable sets.*

PROOF: It suffices to show that $\overline{\mathcal{M}}$ is closed under complementation and under countable unions. Suppose B is any element of $\overline{\mathcal{M}}$ then there exists a measurable set A and a subset N of a Borel set M with $\mu(M) = 0$ such that $B = A \cup N$. But then $B^c = A^c \cap N^c = (A^c \cap M^c) \cup (A^c \cap (M \setminus N))$. On the right, the first term is a measurable set while the second term is a subset of the measure zero set M . It follows that $\overline{\mathcal{M}}$ is closed under complementation.

Now suppose $\{B_i\}$ is a countable sequence of sets in $\overline{\mathcal{M}}$. Then each B_i may be written as the union of a measurable set A_i and a subset N_i of a measurable set M_i of measure zero. As the countable union of sets of measure zero also has measure zero, $M = \bigcup_i M_i$ is also a measurable set of measure zero. It follows that $N = \bigcup_i N_i$ is a subset of the measure zero set M and so $\bigcup_i B_i = N \cup \bigcup_i A_i$ is in $\overline{\mathcal{M}}$. Thus, $\overline{\mathcal{M}}$ is closed under countable unions. ►

Suppose B is any element of the completion $\bar{\mathcal{M}}$. Then $B = A \cup N$ where A is a measurable set and N is a subset of a measure zero set. It is natural then to define a set function $\bar{\mu}$ on the family $\bar{\mathcal{M}}$ by setting $\bar{\mu}(B) = \mu(A)$.

We should verify that the definition is consistent as a set in $\bar{\mathcal{M}}$ may be decomposed in terms of measure zero sets in various ways. Suppose $B = A_1 \cup N_1 = A_2 \cup N_2$ where A_1, A_2 are measurable sets in \mathcal{M} and N_1, N_2 are subsets of measurable sets of measure zero. If we remove $A_1 \cap N_1$ from N_1 and $A_2 \cap N_2$ from N_2 we only make the subsets of measure zero sets smaller without changing B and so we may suppose $A_1 \cap N_1 = A_2 \cap N_2 = \emptyset$. As σ -algebras are closed under symmetric differences, $A_1 \Delta A_2$ is itself a measurable set and hence measurable with respect to μ . But, as I will leave to the reader to verify, $A_1 \Delta A_2 = N_1 \Delta N_2$, so that the symmetric difference of A_1 and A_2 is a measurable subset of a set of measure zero. By monotonicity of measure, it follows that $\mu(A_1 \Delta A_2) = 0$, which implies in turn that $\mu(A_1) = \mu(A_2)$. The definition of $\bar{\mu}(B)$ is hence independent of the choice of decomposition.

I will leave to the reader the trite verification that $\bar{\mu}$ is a measure on the σ -algebra $\bar{\mathcal{M}}$. Moreover, if M is any element of $\bar{\mathcal{M}}$ with $\bar{\mu}(M) = 0$ then all its subsets N are in $\bar{\mathcal{M}}$ (why?) and, by monotonicity, satisfy $\bar{\mu}(N) = 0$. The measure $\bar{\mu}$ on the completion $\bar{\mathcal{M}}$ of the family \mathcal{M} of measurable sets is said to be *complete* in view of this property. It is clear that $\bar{\mu}$ coincides with μ on the measurable sets $A \in \mathcal{M}$ themselves. Thus, $\bar{\mu}$ is an *extension* of μ to the family $\bar{\mathcal{M}}$. There is hence no danger of confusion if we drop the “bar” in the notation and hereafter write just μ for both the original measure on the measurable sets \mathcal{M} and its extension to the family $\bar{\mathcal{M}}$. Wherever needed we will hence assume without comment that the measurable space has been completed and, in particular, that any subset of measure zero is measurable (and has measure zero).

The theory of measure takes on a much more concrete hue in the familiar case of the real line and the construction of probability measures from distribution functions suggests a procedure for the construction of general measures. We begin with any increasing, right continuous, real-valued function $x \mapsto V(x)$, not necessarily bounded, and associate a finite-valued set function $V: (a, b] \mapsto V(b) - V(a)$ on the family of bounded half-closed intervals. If the reader now attempts to extend this set function to the Borel sets of the line by following the steps of the proof of Carathéodory’s extension theorem laid out in the final two sections of this chapter she will find that little more is needed other than to strike out the occasional reference to finiteness of measure. The exercise is salutary and the reader who undertakes it can not only bask in a glow of moral superiority but can lay claim to a fundamental understanding of the rôle played by Carathéodory’s concept of outer measure.

THEOREM 3 *Each increasing, right continuous function V , not necessarily bounded, induces a unique σ -finite measure V on the Borel sets which, to each bounded half-closed interval $(a, b]$, assigns finite measure $V(a, b] = V(b) - V(a)$.*

We call V the *distribution* associated with the measure V though it is not, in general, a probability distribution. If V increases unboundedly then the associated measure V attaches infinite measure to the real line.

3 Lebesgue measure, countable sets

The d.f. increasing linearly in the unit interval is of particular interest. It is easy to see that the function U defined by $U(x) = x$ for $0 < x < 1$ (and taking value zero for $x \leq 0$ and value one for $x \geq 1$) is a bona fide d.f. If $0 < a < b < 1$, the corresponding probability measure U satisfies $U(a, b] = b - a$ on the subintervals of the unit interval. We hence identify U with the ordinary notion of length on subintervals and, more generally, Borel subsets of the unit interval. The measure U is more conventionally written λ and is called *Lebesgue measure on the unit interval*. The reader may recall that she has already encountered this measure in Chapter V, albeit restricted in main to the length of finite unions of intervals.

We may extend the idea to length on the real line though the measure is now no longer finite by a consideration of the *identity map* $I: x \mapsto x$: this map induces the set function $I: (a, b] \mapsto b - a$ which maps half-closed intervals to their lengths. The natural extension of this set function to the Borel sets of the line is called *Lebesgue measure on the line* and is conventionally denoted λ instead of I . The special importance of Lebesgue measure makes it worth recasting the extension theorem for this case.

THEOREM 1 *There exists a unique measure λ on the Borel sets of the line which maps each half-closed interval $(a, b]$ to its length, $\lambda(a, b] = b - a$.*

Where necessary we will assume without comment that the family of Borel sets has been extended to the family of *Lebesgue-measurable sets*, usually denoted \mathcal{L} instead of $\overline{\mathcal{B}}$, via the process of completion by incorporating all subsets of Borel sets of Lebesgue measure zero as sketched in the previous section; the measure λ extends accordingly to this larger family. While Lebesgue measure is pervasive, we will not need the theory of general measures in most of the book barring a couple of scattered references in Chapter XIV.

The construction of Lebesgue measure is one of the great achievements of the theory providing as it does a bare-hands extension of the idea of length to complicated sets on the line. Two questions may, however, have crossed the reader's mind: (1) Are there Lebesgue-measurable sets that are not Borel sets? (2) Are there any subsets of the line that are not Lebesgue-measurable? The answer to both questions is "Yes" though our discussions have not been delicate enough to resolve them. If the reader is familiar with set theory and cardinal arithmetic, the first question can be answered by considering the cardinal numbers of the family of Borel sets and of the family of Lebesgue-measurable sets that are the subsets of the Cantor set in Example 3 to follow; Problem 6

sketches the argument. Resolving the second question requires a delicate excursus into number theory; the interested reader may consult Problems 8–10 to see how this is done.

A consequence of the fact that there exist sets that are not Lebesgue-measurable is the unfortunate conclusion that there is no consistent extension of Lebesgue measure to all the subsets of the real line so that the extended set function is still a measure and is invariant to translations (a characteristic of Lebesgue measure). We will not pursue these issues further here as they will take us somewhat far from our main line of investigation. Fortunately for us, the family of Borel sets (and their completion, the Lebesgue-measurable sets) are rich enough to serve all our needs.

Sets whose Lebesgue measure is zero are of a peculiar importance. Suppose x is an isolated point on the real line. Then, for any $\epsilon > 0$, we may situate x in a half-closed interval \mathbb{I} of length ϵ , for instance, the interval $(x - \epsilon, x]$. By monotonicity of Lebesgue measure, $0 \leq \lambda\{x\} \leq \lambda(\mathbb{I}) = \epsilon$. As ϵ may be selected arbitrarily small, it follows that $\lambda\{x\} = 0$ and all isolated points have Lebesgue measure zero. (This is hardly surprising. The reader will recall from elementary calculus that the ordinary Riemann integral of an isolated point is zero.) An immediate consequence is that *the Lebesgue measure of intervals is insensitive to the addition or deletion of boundary points*: $\lambda(a, b] = \lambda(a, b) = \lambda[a, b] = \lambda[a, b] = b - a$.

It follows readily that the Lebesgue measure of any finite collection of points is also identically zero. But much larger sets of measure zero can be exhibited. We recall that a set is *countable* if its elements can be placed into one-to-one correspondence with the natural numbers. Thus, a set is countable if it is either finite or denumerably infinite. For instance, the set of integers is denumerably infinite and hence countable (as witnessed, for example, by the map which takes each positive integer n to the even number $2n$ and each negative integer $-n$ to the odd number $2n - 1$).

THEOREM 2 *Any countable subset of the real line has Lebesgue measure zero.*

PROOF: We could argue this directly using countable additivity from the fact that a single point has measure zero but the underlying covering argument has its charms. Let $\{x_i, i \geq 1\}$ be an enumeration of a countable set \mathbb{J} . Fix any $\epsilon > 0$. We may then situate each x_i in an interval \mathbb{I}_i of length $\epsilon 2^{-i}$. As $\mathbb{J} \subseteq \bigcup_i \mathbb{I}_i$, it follows that $\lambda(\mathbb{J}) \leq \sum_{i=1}^{\infty} \lambda(\mathbb{I}_i) = \epsilon \sum_{i=1}^{\infty} 2^{-i} = \epsilon$. As the positive ϵ may be chosen arbitrarily small, it follows that $\lambda(\mathbb{J}) = 0$. ▶

Thus, the natural numbers and the integers both have Lebesgue measure zero. A topologically more baffling pair of examples is given below.

EXAMPLES: 1) *Lattice points in the plane.* The set of lattice points (m, n) where m and n are positive integers appears to be clearly much “larger” than the natural numbers. Remarkably, however, Figure 1 suggests how one could snake a line (i.e., a map into the natural numbers) through all the lattice points. Explicitly, each ordered pair of positive integers (m, n) may be mapped into the unique

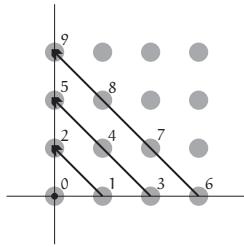


Figure 1: Cantor diagonalisation.

positive integer $m + (m+n)(m+n+1)/2$. This procedure is known as *Cantor diagonalisation*. Lattice points in the other three quadrants may be handled in the same way (or, alternatively, by snaking a line through them in a concentric diamond pattern). It follows that *the set of lattice points in the plane is countable*.

2) *The set \mathbb{Q} of rationals.* Any rational number can be expressed in the form m/n where m is an integer and n a strictly positive integer, and hence can be associated with a pair of integers (m, n) . The rational numbers can hence be put into a one-to-one correspondence with lattice points in the plane. In view of the previous example, it follows that *the rationals are countable!* It follows that the set \mathbb{Q} of rationals has measure zero. From a topological perspective the reader may find this a little troubling. As she is well aware, every non-empty interval of the real line, however small, contains rational points; we express this fact by saying that the rationals are *dense* in the line. Nevertheless their Lebesgue measure (or length) is zero. ►

A consequence of Theorem 2 is that any set of positive measure must be uncountable. As every interval has positive Lebesgue measure (its length!) it follows that *every interval is uncountable, and a fortiori so is the real line*. In light of this, the following classical construction provides a nasty jar to intuition.



Figure 2: Construction of the ternary Cantor set.

EXAMPLE 3) *The Cantor set.* We begin with the closed unit interval $[0, 1]$ and as a first step surgically remove the open interval $(\frac{1}{3}, \frac{2}{3})$ constituting the middle thirds. This leaves two closed subintervals; at the next step we remove

the middle one-third of each of them, that is, we remove the two open intervals $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$. This leaves four closed subintervals; we again remove the middle thirds of each of them, that is, we remove the four open intervals $(\frac{1}{27}, \frac{2}{27})$, $(\frac{7}{27}, \frac{8}{27})$, $(\frac{19}{27}, \frac{20}{27})$, and $(\frac{25}{27}, \frac{26}{27})$. We proceed stage by stage in this fashion; Figure 2 illustrates the process. The open intervals that have been removed through the first five stages are listed in Table 1 and the reader can readily imagine how to extend the process indefinitely.

After stage i , a net total of $1+2+4+\cdots+2^{i-1} = 2^i - 1$ open intervals will have been removed and we enumerate them from left to right as $\mathbb{J}_{i,1}, \mathbb{J}_{i,2}, \dots, \mathbb{J}_{i,2^i-1}$. In this notation, after stage 1 we will have removed the interval $\mathbb{J}_{1,1} = (\frac{1}{3}, \frac{2}{3})$, after stage 2 the three intervals $\mathbb{J}_{2,1} = (\frac{1}{9}, \frac{2}{9}), \mathbb{J}_{2,2} = (\frac{3}{9}, \frac{6}{9}) = (\frac{1}{3}, \frac{2}{3})$, and $\mathbb{J}_{2,3} = (\frac{7}{9}, \frac{8}{9})$, and so on. For each i , the open set $\mathbb{U}_i = \mathbb{J}_{i,1} \cup \dots \cup \mathbb{J}_{i,2^i-1}$ is the set of points that have been removed through stage i in the process.¹ It is clear that $\{\mathbb{U}_i\}$ is an increasing sequence of open sets with limit set $\mathbb{U} = \bigcup_{i=1}^{\infty} \mathbb{U}_i$ the union of all the removed segments. Clearly, \mathbb{U} is an open set. (It is a countable union of open sets.) What is the measure of \mathbb{U} ? As there are exactly 2^{i-1} additional disjoint, open intervals, each of length 3^{-i} , that are excised at a given stage i , through stage i the measure of the set of all intervals excised through that stage satisfies

$$\lambda(\mathbb{U}_i) = \frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots + \frac{2^{i-1}}{3^i} = \frac{1}{3} \left(1 + \frac{2}{3} + \frac{4}{9} + \cdots + \frac{2^{i-1}}{3^{i-1}} \right) = 1 - \left(\frac{2}{3} \right)^i$$

as the expression in the round brackets in the penultimate step sums over a finite geometric series with term $2/3$. By continuity of Lebesgue measure, we hence obtain $\lambda(\mathbb{U}) = \lim_{i \rightarrow \infty} \lambda(\mathbb{U}_i) = 1$. It follows that \mathbb{U} is dense in the unit interval $[0, 1]$. (Else there would exist a subinterval $\mathbb{I} \subset [0, 1]$ of strictly positive measure that contains no points of \mathbb{U} ; but, by additivity of λ , this is incompatible with the measure $\lambda(\mathbb{U}) = 1$.)

The complement of \mathbb{U} with respect to the closed unit interval $[0, 1]$ is called the *Cantor set* $\mathbb{C} = [0, 1] \setminus \mathbb{U}$. As \mathbb{U} is open, \mathbb{C} is closed. Furthermore, $\lambda(\mathbb{C}) = \lambda[0, 1] - \lambda(\mathbb{U}) = 1 - 1 = 0$, and so the *Cantor set* \mathbb{C} has measure zero.

Is \mathbb{C} countable? Let us consider the nature of points in \mathbb{C} . First, it is easy to see that points like $1/3, 8/9$, and $2/27$ are in \mathbb{C} . In general, a point x is in \mathbb{C} if, and only if, it can be expressed in base 3 in the form

$$x = \frac{x_1}{3} + \frac{x_2}{9} + \frac{x_3}{27} + \cdots = \sum_{i=1}^{\infty} \frac{x_i}{3^i}$$

where each x_i takes values 0 or 2 only. (For a number like $1/3$ we have a choice of two representations in base 3, .0̄2 or .1̄0. In such a case we select the former

¹The idea of an open set or a closed set is intuitive in this context; but see Definition I.9.1 for a formal specification.

$(\frac{1}{243}, \frac{2}{243})$
$(\frac{1}{81}, \frac{2}{81})$
$(\frac{7}{243}, \frac{8}{243})$
$(\frac{1}{27}, \frac{2}{27})$
$(\frac{19}{243}, \frac{20}{243})$
$(\frac{7}{81}, \frac{8}{81})$
$(\frac{25}{243}, \frac{26}{243})$
$(\frac{1}{9}, \frac{2}{9})$
$(\frac{55}{243}, \frac{56}{243})$
$(\frac{19}{81}, \frac{20}{81})$
$(\frac{61}{243}, \frac{62}{243})$
$(\frac{7}{27}, \frac{8}{27})$
$(\frac{73}{243}, \frac{74}{243})$
$(\frac{25}{81}, \frac{26}{81})$
$(\frac{79}{243}, \frac{80}{243})$
$(\frac{1}{3}, \frac{2}{3})$
$(\frac{163}{243}, \frac{164}{243})$
$(\frac{55}{81}, \frac{56}{81})$
$(\frac{169}{243}, \frac{170}{243})$
$(\frac{19}{27}, \frac{20}{27})$
$(\frac{181}{243}, \frac{182}{243})$
$(\frac{61}{81}, \frac{62}{81})$
$(\frac{187}{243}, \frac{188}{243})$
$(\frac{7}{9}, \frac{8}{9})$
$(\frac{217}{243}, \frac{218}{243})$
$(\frac{73}{81}, \frac{74}{81})$
$(\frac{223}{243}, \frac{224}{243})$
$(\frac{25}{27}, \frac{26}{27})$
$(\frac{235}{243}, \frac{236}{243})$
$(\frac{79}{81}, \frac{80}{81})$
$(\frac{241}{243}, \frac{242}{243})$

Table 1: The middle-thirds construction.

representation.) If the reader examines the construction carefully she will be able to convince herself why this is true; if she does not succeed, she may wish to take a look at the procedure outlined in Problem XII.12. We map each such $x_i \in \{0, 2\}$ into a binary digit $z_i \in \{0, 1\}$ via $z_i = x_i/2$. Then each x in \mathbb{C} is mapped into a unique number $t = t(x) = \sum_{i=1}^{\infty} \frac{z_i}{2^i}$ in the unit interval. Running through all x in \mathbb{C} will run through all numbers $t = t(x)$ of the above form. But this exhausts the points in the unit interval $[0, 1]$. Thus, we have a one-to-one map from the Cantor set \mathbb{C} onto the unit interval. But, the unit interval has positive Lebesgue measure $\lambda[0, 1] = 1$ and in consequence of Theorem 2 it must be uncountable. *But then this implies that the Cantor set likewise must also be uncountable, yet have measure zero.* ►

Sets of measure zero are fundamental to the theory of measure but have a nuisance value as they cloud arguments and one has to make provision for them with irritating caveats. Some terminology helps finesse the nuisance.

DEFINITION A property of the real numbers that holds everywhere on the line excepting only on a nuisance set of Lebesgue measure zero is said to hold *almost everywhere* (abbreviated *a.e.*).

4 A measure on a ring

To tie up loose ends it only remains to provide a proof of Carathéodory's Extension Theorem of Section 1. We begin with a d.f. F and a set function \mathbf{F} , continuous from above at \emptyset , and defined on the half-closed intervals \mathbb{J} by

$$\mathbf{F}(a, b] = F(b) - F(a), \quad (4.1)$$

the equation reproduced here for ease of reference. How should one proceed to extend the domain of definition of \mathbf{F} from the half-closed intervals to more complex sets? The next step is clearly to meld in additivity by considering finite unions of half-closed intervals. Suppose \mathbb{I} is a subset of the real line that may be expressed as a finite union of disjoint half-closed intervals $\mathbb{I}_1, \dots, \mathbb{I}_n$. Then additivity compels us to define $\mathbf{F}(\mathbb{I})$, if at all, by

$$\mathbf{F}(\mathbb{I}) = \mathbf{F}(\mathbb{I}_1 \cup \dots \cup \mathbb{I}_n) = \mathbf{F}(\mathbb{I}_1) + \dots + \mathbf{F}(\mathbb{I}_n). \quad (4.2)$$

There is a minor subtlety here as the set \mathbb{I} may be partitioned in different ways into a finite collection of half-closed intervals and we have to show that the value of the sum on the right is independent of the partitioning. That this is the case is seen by considering the situation when \mathbb{I} is itself a half-closed interval, say, $(a, b]$. We may then order the selected partitioning so that $a = a_1 < b_1 = a_2 < b_2 = a_3 < \dots < b_{n-1} = a_n < b_n = b$. The right-hand side of the equation is then seen to telescope to the value $F(b_n) - F(a_1) = F(b) - F(a) = \mathbf{F}(a, b]$ which is independent of the specific partitioning chosen. If now, more generally, \mathbb{I} is a union of disjoint half-closed intervals $\mathbb{J}_1, \dots, \mathbb{J}_m$ then we may group the intervals \mathbb{I}_i into m disjoint subgroups with each subgroup covering one of the intervals \mathbb{J}_j . As the value of each $\mathbf{F}(\mathbb{J}_j)$ is independent of the way the relevant

subgroup of intervals \mathbb{I}_i partitions it, the value of the sum of the terms $F(\mathbb{J}_j)$ is also invariant to the partitioning.

As we may express any finite union of half-closed intervals as a finite union of *disjoint* half-closed intervals, in this fashion we have now extended the definition of the set function F to the family $R(\mathcal{J})$ of subsets of the line that may be expressed as a finite union of half-closed intervals. As the set difference of two half-closed intervals, $(a, b] \setminus (c, d]$ is either empty or another half-closed interval, it follows that if \mathbb{I}, \mathbb{I}' are elements of $R(\mathcal{J})$ then so is $\mathbb{I} \setminus \mathbb{I}'$; and, of course, it is palpable that $\mathbb{I} \cup \mathbb{I}'$ is also a member of $R(\mathcal{J})$ as \mathbb{I} and \mathbb{I}' are both finite unions of half-closed intervals so that their union is as well. These properties are captured in a suggestive language: we say that $R(\mathcal{J})$ is *closed under set differences and unions*. In the algebraist's terminology, any family of sets closed under set differences and unions is called a *ring*. The family $R(\mathcal{J})$ is the smallest ring containing the half-closed intervals; it is called the ring *generated by \mathcal{J}* .

Accordingly, the relations (4.1,4.2) uniquely define the set function F on the ring $R(\mathcal{J})$. What are its properties?

THEOREM 1 *The set function F restricted to the ring $R(\mathcal{J})$ is positive and countably additive, and satisfies $F(\emptyset) = 0$.*

The theorem is equivalent to the statement that *the relations (4.1,4.2) define a positive measure on the ring $R(\mathcal{J})$* .

We break the proof up into several pieces which we present in a sequence of lemmas which, together, will complete the proof of the theorem and a little more besides.

LEMMA 1 *The set function F on $R(\mathcal{J})$ is bounded between 0 and 1 and satisfies $F(\emptyset) = 0$.*

PROOF: The positivity of F is trite from the defining relations and as F is continuous from above at \emptyset , we have $F(\emptyset) = 0$. Now suppose \mathbb{I} is any element of $R(\mathcal{J})$. Then we may express \mathbb{I} as a finite union of disjoint half-closed intervals, say, $(a_1, b_1], \dots, (a_n, b_n]$ where we may order the intervals so that $a_1 < b_1 \leq a_2 < b_2 \leq \dots \leq a_n < b_n$. We hence obtain

$$\begin{aligned} F(\mathbb{I}) &= [F(b_1) - F(a_1)] + [F(b_2) - F(a_2)] + \dots + [F(b_n) - F(a_n)] \\ &= F(b_n) - \sum_{i=2}^n [F(a_i) - F(b_{i-1})] - F(a_1) \leq F(b_n) - F(a_1) \leq F(\infty) - F(-\infty) = 1 \end{aligned}$$

as the d.f. F is increasing. ▶

What is not obvious and deserves a detailed proof is the assertion that F is in fact countably additive on $R(\mathcal{J})$. In other words: *if $\{\mathbb{I}_i\}$ is a countable sequence of mutually disjoint elements of $R(\mathcal{J})$ whose union $\mathbb{I} = \bigcup_i \mathbb{I}_i$ is another element of $R(\mathcal{J})$, then $F(\mathbb{I}) = \sum_i F(\mathbb{I}_i)$* . To clear the decks for a proof of the countable additivity property we start with some simple preliminary observations.

LEMMA 2 *Suppose $\mathbb{I}_1, \dots, \mathbb{I}_n$ is a finite collection of disjoint sets in $R(\mathcal{J})$. Then*

$$F(\mathbb{I}_1 \cup \dots \cup \mathbb{I}_n) = F(\mathbb{I}_1) + \dots + F(\mathbb{I}_n).$$

PROOF: It will suffice to show the result for $n = 2$. If \mathbb{I}, \mathbb{J} are disjoint sets in $R(\mathcal{I})$ then we may write them in the form $\mathbb{I} = \bigcup_{i=1}^n \mathbb{I}_i$ and $\mathbb{J} = \bigcup_{j=1}^m \mathbb{J}_j$ where $\mathbb{I}_1, \dots, \mathbb{I}_n, \mathbb{J}_1, \dots, \mathbb{J}_m$ are mutually disjoint half-closed intervals. But then $\mathbb{I} \cup \mathbb{J}$ is the union of disjoint intervals whence

$$F(\mathbb{I} \cup \mathbb{J}) = \sum_{i=1}^n F(\mathbb{I}_i) + \sum_{j=1}^m F(\mathbb{J}_j) = F(\mathbb{I}) + F(\mathbb{J}).$$

The general result follows by induction. ►

As is usual, positivity and additivity confer monotonicity. The proof is virtually the same as for general probability measures.

LEMMA 3 *If \mathbb{I}, \mathbb{I}' are in $R(\mathcal{I})$ and $\mathbb{I} \subseteq \mathbb{I}'$ then $F(\mathbb{I}) \leq F(\mathbb{I}')$.*

PROOF: If \mathbb{I} and \mathbb{I}' are sets in $R(\mathcal{I})$ then, indeed, $\mathbb{J} = \mathbb{I}' \setminus \mathbb{I}$ is also in $R(\mathcal{I})$ as it is closed under set differences. But then $\mathbb{I}' = \mathbb{I} \cup \mathbb{J}$ is the union of disjoint elements of $R(\mathcal{I})$. By additivity it follows that $F(\mathbb{I}') = F(\mathbb{I}) + F(\mathbb{J}) \geq F(\mathbb{I})$, the final step following by positivity of F . ►

The next result is useful in generating bounds when a union is not necessarily of disjoint elements. The reader will recognise a form of Boole's inequality though I will reproduce the proof for this particular context.

LEMMA 4 *Suppose $\mathbb{I}_1, \dots, \mathbb{I}_n$ are arbitrary elements of $R(\mathcal{I})$ (not necessarily disjoint). Then*

$$F(\mathbb{I}_1 \cup \dots \cup \mathbb{I}_n) \leq F(\mathbb{I}_1) + \dots + F(\mathbb{I}_n).$$

PROOF: It will again suffice to prove the result for $n = 2$. Suppose \mathbb{I} and \mathbb{J} are elements of $R(\mathcal{I})$. Then $\mathbb{I} \cup \mathbb{J} = \mathbb{I} \cup (\mathbb{J} \setminus \mathbb{I})$ and the expression on the right is the union of two disjoint sets in $R(\mathcal{I})$. By additivity, $F(\mathbb{I} \cup \mathbb{J}) = F(\mathbb{I}) + F(\mathbb{J} \setminus \mathbb{I}) \leq F(\mathbb{I}) + F(\mathbb{J})$, the final step following by monotonicity as $\mathbb{J} \setminus \mathbb{I} \subseteq \mathbb{J}$. The general result follows by induction. ►

With the timber cleared we are now ready to show that F is countably additive on $R(\mathcal{I})$. We will actually show a little more.

LEMMA 5 *Suppose $\{\mathbb{I}_i\}$ is a countable sequence of elements of the ring $R(\mathcal{I})$ which cover an element \mathbb{I} in $R(\mathcal{I})$. Then $F(\mathbb{I}) \leq \sum_i F(\mathbb{I}_i)$. If the sets \mathbb{I}_i are mutually disjoint and $\bigcup_i \mathbb{I}_i$ is also an element of $R(\mathcal{I})$ then $F(\bigcup_i \mathbb{I}_i) = \sum_i F(\mathbb{I}_i)$.*

PROOF: We prepare the way for showing countable additivity on the ring $R(\mathcal{I})$ (which is the second of the two assertions in the lemma) by first showing that F is countably additive on the family \mathcal{I} . This is the crux of the proof.

Suppose $\{\mathbb{I}_i, i \geq 1\}$ is a disjoint sequence of half-closed intervals and suppose $\bigcup_{i=1}^{\infty} \mathbb{I}_i = \mathbb{I}$ where \mathbb{I} is another half-closed interval. To show that $\sum_i F(\mathbb{I}_i) = F(\mathbb{I})$ it will suffice to show that (1) $\sum_i F(\mathbb{I}_i) \leq F(\mathbb{I})$, and (2) $\sum_i F(\mathbb{I}_i) \geq F(\mathbb{I})$.

The first inequality is the easy part. It is clear that $\bigcup_{i=1}^n \mathbb{I}_i \subseteq \mathbb{I}$ for each n . By monotonicity it follows that $F(\mathbb{I}) \geq F(\mathbb{I}_1 \cup \dots \cup \mathbb{I}_n) = F(\mathbb{I}_1) + \dots + F(\mathbb{I}_n)$, the final step justified by (finite) additivity. As the inequality holds for each value of n , taking the

limit as $n \rightarrow \infty$ of both sides shows that $F(\mathbb{I}) \geq \sum_{i=1}^{\infty} F(\mathbb{I}_i)$, completing the first part of what is to be shown.

To show that the inequality also goes the other way requires more effort. Suppose $\mathbb{I} = (a, b]$. A little thought shows that most of the effort to cover \mathbb{I} with the intervals $\{\mathbb{I}_i\}$ must be expended in the immediate vicinity of the point a . The right continuity of the d.f. F will help wrap up the argument.

One must either be a very sophisticated or a very naïve mathematician to accept the previous paragraph as sufficient justification though, in fact, it contains the essence of the argument. In detail the demonstration goes as follows. Fix any $\epsilon > 0$. As the d.f. F is right continuous, we may select $0 < \delta < b - a$ small enough so that $F(a, a + \delta] < \epsilon$. Suppose $\mathbb{I}_i = (a_i, b_i]$. Again by right continuity of F , for each $i \geq 1$ we may select $\delta_i > 0$ sufficiently small so that $F(b_i, b_i + \delta_i) < \epsilon/2^i$. By assumption, the collection of half-closed intervals $\{(a_i, b_i], i \geq 1\}$ forms a cover of $(a, b]$. As $(a_i, b_i] \subset (a_i, b_i + \delta_i)$ for each i and $[a + \delta, b] \subset (a, b]$, it follows *a fortiori* that the collection of open intervals $\{(a_i, b_i + \delta_i), i \geq 1\}$ forms an open cover of the closed interval $[a + \delta, b]$. By the Heine-Borel theorem of Section XXI.1 in the Appendix there exists a finite subcover $\{(a_1, b_1 + \delta_1), \dots, (a_n, b_n + \delta_n)\}$ of $[a + \delta, b]$ for some sufficiently large n whence

$$(a + \delta, b] \subset [a + \delta, b] \subseteq \bigcup_{i=1}^n (a_i, b_i + \delta_i) \subset \bigcup_{i=1}^n (a_i, b_i + \delta_i).$$

It follows that

$$\begin{aligned} F(a + \delta, b] &\stackrel{(i)}{\leq} F\left(\bigcup_{i=1}^n (a_i, b_i + \delta_i)\right) \stackrel{(ii)}{\leq} \sum_{i=1}^n F(a_i, b_i + \delta_i) \stackrel{(iii)}{\leq} \sum_{i=1}^{\infty} F(a_i, b_i + \delta_i) \\ &\stackrel{(iv)}{=} \sum_{i=1}^{\infty} [F(a_i, b_i) + F(b_i, b_i + \delta_i)] \stackrel{(v)}{\leq} \sum_{i=1}^{\infty} F(a_i, b_i) + \epsilon \sum_{i=1}^{\infty} 2^{-i} \stackrel{(vi)}{=} \sum_{i=1}^{\infty} F(a_i, b_i] + \epsilon, \end{aligned}$$

where (i) follows by monotonicity, (ii) by subadditivity, (iii) by positivity, (iv) by additivity, (v) by right continuity and the choice of the δ_i , and (vi) as the geometric series sums to 1. On the other hand, by additivity,

$$F(a + \delta, b] = F(a, b] - F(a, a + \delta] \geq F(a, b] - \epsilon$$

by choice of the sufficiently small, positive δ . Accordingly, $F(a, b] \leq \sum_{i=1}^{\infty} F(a_i, b_i] + 2\epsilon$ and, as the positive ϵ may be chosen arbitrarily small, it follows that

$$F(\mathbb{I}) = F(a, b] \leq \sum_{i=1}^{\infty} F(a_i, b_i] = \sum_{i=1}^{\infty} F(\mathbb{I}_i)$$

which is the second part of what we set out to show.

We've hence shown that F is countably additive on \mathcal{J} . We now extend the argument to show that indeed F is countably additive over the ring generated by \mathcal{J} as well. Suppose $\{\mathbb{I}_i\}$ is a countable collection of disjoint sets in $R(\mathcal{J})$ and suppose their union \mathbb{I} is also in $R(\mathcal{J})$. Then each \mathbb{I}_i may be represented as the union of a finite number of disjoint half-closed intervals, say $\mathbb{I}_{i1}, \dots, \mathbb{I}_{in_i}$ and, by (finite) additivity, $F(\mathbb{I}_i) = \sum_j F(\mathbb{I}_{ij})$. Suppose first that \mathbb{I} itself is a half-closed interval in \mathcal{J} . Then since the collection of all \mathbb{I}_i is countable and disjoint and $\mathbb{I} = \bigcup_i \bigcup_j \mathbb{I}_{ij}$, by countable additivity of F over \mathcal{J} it follows that $F(\mathbb{I}) = \sum_i \sum_j F(\mathbb{I}_{ij}) = \sum_i F(\mathbb{I}_i)$. Now suppose in general that \mathbb{I} is a finite, disjoint union of sets in \mathcal{J} , $\mathbb{I} = \mathbb{J}_1 \cup \dots \cup \mathbb{J}_m$. As $\{\mathbb{I}_i\}$ covers \mathbb{I} , it follows that, for each k , the family

of sets $\{\mathbb{I}_i \cap \mathbb{J}_k\}$ is a cover of \mathbb{J}_k ; the reader should bear in mind that, for each i , $\mathbb{I}_i \cap \mathbb{J}_k$ is a finite union of half-closed intervals and is hence itself an element of the ring $R(\mathcal{J})$. It follows that

$$F(\mathbb{I}) \stackrel{(vii)}{=} \sum_k F(\mathbb{J}_k) \stackrel{(viii)}{=} \sum_k \sum_i F(\mathbb{I}_i \cap \mathbb{J}_k) \stackrel{(ix)}{=} \sum_i \sum_k F(\mathbb{I}_i \cap \mathbb{J}_k) \stackrel{(x)}{=} \sum_i F(\mathbb{I}_i),$$

where (vii) is validated by the finite additivity of F over $R(\mathcal{J})$, (viii) follows by the just-proved specialisation of the desired result to half-closed intervals as the half-closed interval \mathbb{J}_k may be represented as the disjoint countable union of the elements $\mathbb{I}_i \cap \mathbb{J}_k$, each of which is in the ring $R(\mathcal{J})$, (ix) holds as one may sum a series in any order if the summands are positive, and (x) again follows by the finite additivity of F over $R(\mathcal{J})$ because the set \mathbb{I}_i in $R(\mathcal{J})$ may be represented as the disjoint finite union of the elements $\mathbb{I}_i \cap \mathbb{J}_1, \dots, \mathbb{I}_i \cap \mathbb{J}_m$ in $R(\mathcal{J})$. We've hence established that, as advertised, the set function F is countably additive over the ring $R(\mathcal{J})$ generated by the half-closed intervals \mathcal{J} .

To finish up, we need to demonstrate the validity of the first assertion in the lemma. Suppose $\{\mathbb{I}_i\}$ is a cover of \mathbb{I} , all sets in $R(\mathcal{J})$ and the \mathbb{I}_i not necessarily disjoint. For each i , write $\mathbb{J}_i = \mathbb{I}_i \cap \mathbb{I}$. As rings are closed under set differences, for each i , $\mathbb{I}_i \setminus \mathbb{I}$ is also an element of $R(\mathcal{J})$, whence so is $\mathbb{J}_i = \mathbb{I}_i \cap \mathbb{I} = \mathbb{I}_i \setminus (\mathbb{I}_i \setminus \mathbb{I})$. As $\mathbb{I} \subseteq \bigcup_i \mathbb{I}_i$, it follows that the family $\{\mathbb{J}_i\}$ of smaller sets in $R(\mathcal{J})$ also forms a countable cover of \mathbb{I} . Now, in general, the sets \mathbb{J}_i are not disjoint. We can form an equivalent disjoint cover, however, by setting $\mathbb{K}_1 = \mathbb{J}_1$ and, for $i \geq 2$, $\mathbb{K}_i = \mathbb{J}_i \setminus (\mathbb{K}_1 \cup \dots \cup \mathbb{K}_{i-1})$. Then $\bigcup_i \mathbb{K}_i = \bigcup_i \mathbb{J}_i = \mathbb{I}$. As the family $\{\mathbb{K}_i\}$ is a sequence of mutually disjoint sets in $R(\mathcal{J})$ whose union is \mathbb{I} , by countable additivity, $F(\mathbb{I}) = \sum_i F(\mathbb{K}_i) \leq \sum_i F(\mathbb{J}_i) \leq \sum_i F(\mathbb{I}_i)$, the successive inequalities validated by monotonicity of F over $R(\mathcal{J})$ as, for each i , we have the obvious set inclusions $\mathbb{K}_i \subseteq \mathbb{J}_i \subseteq \mathbb{I}_i$. ▶

The sequence of lemmas establish that F is a positive measure on $R(\mathcal{J})$ as was to be shown.



5 From measure to outer measure, and back

We are still some distance from where we wish to go but we make progress. An *algebra* is simply a ring that contains the entire set (in our case \mathbb{R}) as one of its elements. Thus, equivalently, an algebra is closed under complementation and finite unions and includes the entire set. As the real line cannot be expressed as a finite union of bounded, half-closed intervals, it is clear in the present instance that the ring $R(\mathcal{J})$ does not contain \mathbb{R} and is hence not an algebra. We will need to fatten it up a little to make it not just an algebra but a σ -algebra closed under complementation and countable unions.

One can extend the line of thought leading up to the specification of the measure F on the ring $R(\mathcal{J})$ to countable unions of intervals but then one appears to hit an impasse. It is a sad fact but true that a general Borel set need not be expressible as a countable union of intervals and it is now not nearly so clear how one should proceed from this point; or indeed whether there is an unambiguous extension of the set function F from the ring $R(\mathcal{J})$ to more complex sets.

One may naïvely hope, for instance, to extend F to a probability measure on all subsets of the line. But it proves impossible to extend F consistently in such a manner.

The right approach was found by Carathéodory. Instead of attempting to design the measure F itself, we proceed indirectly. In lieu of a measure, we first specify an envelope (or *outer measure*) F^* on all subsets of the line and which agrees with F on the ring generated by the intervals; this, as we shall see, can be done unambiguously and simply. The outer measure fails to be a measure proper because in general it is only countably subadditive and not additive. However, it will turn out to be countably additive on the sets that matter—the Borel sets of the line, and indeed a little bit more. The restriction of the outer measure to the Borel sets then specifies the desired probability measure F on the Borel σ -algebra of subsets of the line.

DEFINITION 1 The *outer measure* induced by the d.f. F is the set function F^* defined on all subsets A of the real line via $F^*(A) = \inf \sum_i F(I_i)$ where the infimum is over the set of values obtained for the sum on the right when $\{I_i\}$ is allowed to range over all countable covers of A with the individual sets I_i all constrained to be elements of the ring $R(\mathcal{J})$.

As each element of $R(\mathcal{J})$ is a finite union of intervals, the countable union of elements of $R(\mathcal{J})$ may be expressed as a countable union of intervals. Accordingly, we may just as well have restricted the elements I_i in the definition to the family of half-closed intervals \mathcal{J} . In its stated form the definition gives us a little more flexibility in arguments.

The outer measure specifies an envelope for the size of a given set by considering the sum of the measures of a cover of the set. Prudence dictates that we consider the smallest (or, more precisely, the greatest lower bound) of these values to get at the slimmest envelope containing the given set. The reader should remark that there is no danger of the set over which the infimum is to be taken being empty. The countable sequence of half-closed intervals $(i-1, i]$ as i ranges over all integers covers \mathbb{R} and hence also any subset of \mathbb{R} .

We should immediately take stock of the properties of the outer measure to verify that it does indeed have the desired features.

THEOREM 1 *The outer measure F^* agrees with F on the sets of the ring $R(\mathcal{J})$. Moreover, F^* is positive, bounded, monotone, and countably subadditive on the subsets of the real line, and satisfies $F^*(\emptyset) = 0$ and $F^*(\mathbb{R}) = 1$.*

As in the previous section, I will present a collection of partial results in a succession of lemmas that together will establish what is claimed in the theorem, and more besides. We first begin by showing that the outer measure is indeed an extension of the measure on the ring constructed in the previous section.

LEMMA 1 *The outer measure F^* agrees with F on the sets of the ring $R(\mathcal{J})$, that is to say, $F^*(I) = F(I)$ for all I in $R(\mathcal{J})$. In other words, the outer measure F^* constitutes an extension of the measure F on the ring $R(\mathcal{J})$ to a set function on all the subsets of the line.*

PROOF: Fix any element I of the ring $R(\mathcal{J})$. Clearly, I is covered by the sequence of sets $I, \emptyset, \emptyset, \dots$, all of which are trivially sets of $R(\mathcal{J})$. It follows that

$$F^*(I) \leq F(I) + F(\emptyset) + F(\emptyset) + \dots = F(I).$$

On the other hand, if $\{I_i\}$ is any countable collection of sets in $R(\mathcal{J})$ that forms a cover of I then, by Lemma 4.5, $F(I) \leq \sum_i F(I_i)$. Since this is valid for all covers $\{I_i\}$ of I , the

inequality is valid for the infimum of the right-hand side over all covers; or, what is the same thing, $F(\mathbb{I}) \leq F^*(\mathbb{I})$. It follows that, indeed, $F^*(\mathbb{I}) = F(\mathbb{I})$ for all \mathbb{I} in $R(\mathcal{J})$. ►

We now establish the properties of the outer measure to mirror those of the measure on the ring. We begin with monotonicity.

LEMMA 2 *If \mathbb{A} and \mathbb{B} are subsets of the real line with $\mathbb{A} \subseteq \mathbb{B}$ then $F^*(\mathbb{A}) \leq F^*(\mathbb{B})$.*

PROOF: If $\mathbb{A} \subseteq \mathbb{B}$ then every cover of \mathbb{B} is also a cover of \mathbb{A} . As $F^*(\mathbb{A})$ is obtained as the infimum over a larger set we must have $F^*(\mathbb{A}) \leq F^*(\mathbb{B})$. ►

Positivity is trite; it is worth noting, however, that the outer measure properly normalises the real line to outer measure 1.

LEMMA 3 *The outer measure F^* is positive and bounded with $0 \leq F^*(\mathbb{A}) \leq 1$ for all subsets \mathbb{A} of the line. Furthermore, $F^*(\emptyset) = 0$ and $F^*(\mathbb{R}) = 1$.*

PROOF: It is clear that F^* is positive over its entire domain as F is positive and the greatest lower bound of a set of positive numbers cannot be < 0 . Furthermore, $F^*(\emptyset) = F(\emptyset) = 0$ as F^* agrees with F over the ring generated by \mathcal{J} and, in particular, on the empty set. Finally, the entire real line \mathbb{R} is covered by the countable collection of half-open intervals $(i-1, i]$ with i ranging over all the integers. Consequently,

$$F^*(\mathbb{R}) \leq \sum_i F(i-1, i] = \sum_i [F(i) - F(i-1)] = F(\infty) - F(-\infty) = 1$$

as the sum telescopes. On the other hand, $(a, b] \subset \mathbb{R}$ for all half-closed intervals $(a, b]$. It follows by the monotonicity of F^* on the subsets of \mathbb{R} that

$$F^*(\mathbb{R}) \geq F^*(a, b] = F(a, b] = F(b) - F(a)$$

for every choice of a and b as F^* and F agree on the ring $R(\mathcal{J})$ and, *a fortiori* on the half-closed intervals \mathcal{J} . Allowing a to tend to $-\infty$ and b to $+\infty$ on both sides, we obtain

$$F^*(\mathbb{R}) \geq F(\infty) - F(-\infty) = 1 - 0 = 1.$$

It follows that $F^*(\mathbb{R}) = 1$ identically. By the monotonicity of F^* , for any subset \mathbb{A} of \mathbb{R} we must have $F^*(\emptyset) \leq F^*(\mathbb{A}) \leq F^*(\mathbb{R})$, completing the proof. ►

And, finally, we come to the key countable subadditivity property. With the following lemma in hand we will have completed the proof of Theorem 1.

LEMMA 4 *Suppose \mathbb{A} is any subset of \mathbb{R} and $\{\mathbb{A}_i, i \geq 1\}$ is any family of subsets of \mathbb{R} that forms a countable cover of \mathbb{A} . Then $F^*(\mathbb{A}) \leq \sum_i F^*(\mathbb{A}_i)$.*

PROOF: Fix any $\epsilon > 0$. By the definition of infimum, for each \mathbb{A}_i there exists a countable collection $\{\mathbb{I}_{ij}, j \geq 1\}$ of elements of $R(\mathcal{J})$ that cover \mathbb{A}_i and such that $F^*(\mathbb{A}_i) \geq \sum_j F(\mathbb{I}_{ij}) - \epsilon/2^i$. But then the entire collection of ring elements $\{\mathbb{I}_{ij}\}$ with both i and j varying over all possibilities constitutes a countable cover of \mathbb{A} . It follows that

$$F^*(\mathbb{A}) \leq \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} F(\mathbb{I}_{ij}) \leq \sum_{i=1}^{\infty} \left[F^*(\mathbb{A}_i) + \frac{\epsilon}{2^i} \right] = \sum_{i=1}^{\infty} F^*(\mathbb{A}_i) + \epsilon \sum_{i=1}^{\infty} \frac{1}{2^i} = \sum_{i=1}^{\infty} F^*(\mathbb{A}_i) + \epsilon$$

as the geometric series again sums to one. As the inequality is valid for all $\epsilon > 0$, we must have $F^*(A) \leq \sum_i F^*(A_i)$ as asserted. ▶

We cannot in general ensure that the outer measure F^* is countably additive over the family of all subsets of the real line—only that it is countably subadditive. For what family of subsets of \mathbb{R} then can we hope to obtain countable additivity? The key is provided by a curious additivity property of outer measure with respect to the elements of the ring $R(\mathcal{I})$.

LEMMA 5 Suppose $M \in R(\mathcal{I})$ is any element of the ring generated by the half-closed intervals and A is any subset of the real line. Then

$$F^*(A) = F^*(A \cap M) + F^*(A \cap M^c). \quad (5.1)$$

PROOF: As is typical in proofs of this type, we will show that the identity (5.1) holds by establishing an inequality in both directions. One side is easy.

It is clear that we may decompose A as the disjoint union $A = (A \cap M) \cup (A \cap M^c)$. By the established subadditivity of F^* it follows that

$$F^*(A) \leq F^*(A \cap M) + F^*(A \cap M^c).$$

To prove that the inequality also holds in reverse, fix any $\epsilon > 0$ and pick a cover $\{\mathbb{I}_i\}$ of A with each $\mathbb{I}_i \in R(\mathcal{I})$ and such that

$$F^*(A) \geq \sum_i F(\mathbb{I}_i) - \epsilon.$$

(The fact that we can do this follows by the definition of the infimum of a set.) But we may decompose \mathbb{I}_i into the disjoint union of the elements $\mathbb{I}_i \cap M$ and $\mathbb{I}_i \setminus M (= \mathbb{I}_i \cap M^c)$, both of which are elements of $R(\mathcal{I})$. As F is additive over $R(\mathcal{I})$, it follows that $F(\mathbb{I}_i) = F(\mathbb{I}_i \cap M) + F(\mathbb{I}_i \cap M^c)$. It follows that

$$F^*(A) \geq \sum_i F(\mathbb{I}_i \cap M) + \sum_i F(\mathbb{I}_i \cap M^c) - \epsilon.$$

But $\{\mathbb{I}_i \cap M\}$ and $\{\mathbb{I}_i \cap M^c\}$ are covers of $A \cap M$ and $A \cap M^c$, respectively, whence $F^*(A \cap M) \leq \sum_i F(\mathbb{I}_i \cap M)$ and $F^*(A \cap M^c) \leq \sum_i F(\mathbb{I}_i \cap M^c)$. It follows that

$$F^*(A) \geq F^*(A \cap M) + F^*(A \cap M^c) - \epsilon.$$

As this is valid for each $\epsilon > 0$, we must have

$$F^*(A) \geq F^*(A \cap M) + F^*(A \cap M^c).$$

The two pieces of the inequality establish the claimed result. ▶

As we've seen that the outer measure F^* is an extension of the measure F on the ring $R(\mathcal{I})$, it is clear that outer measure is countably additive over at least the family of sets $R(\mathcal{I})$. In view of the additive identity (5.1) we may wonder whether countable additivity is preserved for the (potentially larger) family of subsets of the line for which (5.1) holds. At any rate this is worth investigating.

Let $\overline{R(\mathcal{I})}$ be the family of subsets M of the real line for which the additive identity (5.1) holds for every subset A of the line. Clearly, $\overline{R(\mathcal{I})}$ contains the ring $R(\mathcal{I})$, hence also the family \mathcal{I} of half-closed intervals. What more can be said?

THEOREM 2 *The family $\overline{R(\mathcal{I})}$ of subsets of the real line is closed under complementations and countable unions and contains \mathbb{R} as an element as well as all the half-closed intervals. In other words, $\overline{R(\mathcal{I})}$ is a σ -algebra containing the family \mathcal{I} of half-closed intervals.*

PROOF: It is clear that all the half-closed intervals, and indeed all elements of the ring $R(\mathcal{I})$ are contained in $\overline{R(\mathcal{I})}$. As $F^*(\emptyset) = 0$, it follows that, for any subset A of the reals, $F^*(A \cap \mathbb{R}) + F^*(A \cap \mathbb{R}^c) = F^*(A) + F^*(\emptyset) = F^*(A)$, so that the set of real numbers \mathbb{R} is also an element of $\overline{R(\mathcal{I})}$.

Closure under complementation is trite by the defining equation (5.1): if M is in $\overline{R(\mathcal{I})}$ then so is M^c . This only leaves us to verify that $\overline{R(\mathcal{I})}$ is closed under countable unions. We begin with finite unions.

Suppose M and N are elements of $\overline{R(\mathcal{I})}$ and let A be any subset of \mathbb{R} . By two usages of (5.1) we obtain

$$\begin{aligned} F^*(A) &= F^*(A \cap M) + F^*(A \cap M^c) \\ &= F^*(A \cap M \cap N) + F^*(A \cap M \cap N^c) + F^*(A \cap M^c \cap N) + F^*(A \cap M^c \cap N^c). \end{aligned} \quad (5.2)$$

The argument of the last term on the right may be written in the form $A \cap M^c \cap N^c = A \cap (M \cup N)^c$. Accordingly, if in (5.2) we replace A by $A \cap (M \cup N)$, the last term vanishes and as $M \subseteq M \cup N$ and $N \subseteq M \cup N$, we are left with the identity

$$F^*(A \cap (M \cup N)) = F^*(A \cap M \cap N) + F^*(A \cap M \cap N^c) + F^*(A \cap M^c \cap N). \quad (5.3)$$

It follows that

$$F^*(A) = F^*(A \cap (M \cup N)) + F^*(A \cap (M \cup N)^c)$$

whence $M \cup N$ is in $\overline{R(\mathcal{I})}$. That $\overline{R(\mathcal{I})}$ is closed under finite unions follows readily by induction: if M_1, \dots, M_n are elements of $\overline{R(\mathcal{I})}$ then so is $M_1 \cup \dots \cup M_n$.

We now proceed to countable disjoint unions. Suppose first that M and N are disjoint sets in $\overline{R(\mathcal{I})}$. Then $M \subseteq N^c$ and $N \subseteq M^c$ and the identity (5.3) simplifies to

$$F^*(A \cap (M \cup N)) = F^*(A \cap M) + F^*(A \cap N).$$

Suppose now that $\{M_i\}$ is a countable sequence of disjoint sets in $\overline{R(\mathcal{I})}$. By induction, we obtain

$$F^*(A \cap (M_1 \cup \dots \cup M_n)) = F^*(A \cap M_1) + \dots + F^*(A \cap M_n).$$

For each $n \geq 1$, set $N_n = M_1 \cup \dots \cup M_n$, and let $M = \bigcup_{i=1}^{\infty} M_i$. Then N_n is in $\overline{R(\mathcal{I})}$ (as $\overline{R(\mathcal{I})}$ is closed under finite unions). It follows that, for any subset A of the reals,

$$\begin{aligned} F^*(A) &= F^*(A \cap N_n) + F^*(A \cap N_n^c) \\ &= \sum_{i=1}^n F^*(A \cap M_i) + F^*(A \cap N_n^c) \geq \sum_{i=1}^n F^*(A \cap M_i) + F^*(A \cap M^c), \end{aligned}$$

the final step following by monotonicity as $N_n \subseteq M$. Taking the limit as $n \rightarrow \infty$, we hence obtain

$$F^*(A) \geq \sum_{i=1}^{\infty} F^*(A \cap M_i) + F^*(A \cap M^c) \geq F^*(A \cap M) + F^*(A \cap M^c) \quad (5.4)$$

as F^* is subadditive and \mathbb{M} is the union of the sets $\{\mathbb{M}_i\}$ and hence *a fortiori* $\mathbb{M} \subseteq \bigcup_i \mathbb{M}_i$. But \mathbb{A} is the union of the sets $\mathbb{A} \cap \mathbb{M}$ and $\mathbb{A} \cap \mathbb{M}^c$ so that \mathbb{A} is covered by these two sets. By subadditivity of F^* it follows again that

$$F^*(\mathbb{A}) \leq F^*(\mathbb{A} \cap \mathbb{M}) + F^*(\mathbb{A} \cap \mathbb{M}^c). \quad (5.5)$$

As the inequality holds in both directions it must follow indeed that

$$F^*(\mathbb{A}) = F^*(\mathbb{A} \cap \mathbb{M}) + F^*(\mathbb{A} \cap \mathbb{M}^c)$$

whence \mathbb{M} is contained in $\overline{R(\mathcal{I})}$.

We've hence shown that $\overline{R(\mathcal{I})}$ is closed under countable disjoint unions. But any countable union of sets may be replaced by a countable disjoint union via a, by now, familiar approach. Suppose $\{\mathbb{M}_i\}$ is a countable sequence of sets in $\overline{R(\mathcal{I})}$ and let $\mathbb{M} = \bigcup_i \mathbb{M}_i$. Then the family $\{\mathbb{N}_i\}$ of sets defined by $\mathbb{N}_1 = \mathbb{M}_1$ and, for $i \geq 2$, $\mathbb{N}_i = \mathbb{M}_i \setminus (\mathbb{N}_1 \cup \dots \cup \mathbb{N}_{i-1})$ is disjoint, with each $\mathbb{N}_i \in \overline{R(\mathcal{I})}$, and with $\mathbb{M} = \bigcup_i \mathbb{N}_i = \bigcup_i \mathbb{M}_i$. We may hence replace $\{\mathbb{M}_i\}$ by the disjoint family $\{\mathbb{N}_i\}$ without affecting the union \mathbb{M} . And thus, $\overline{R(\mathcal{I})}$ is closed under arbitrary countable unions. ▶

As $\overline{R(\mathcal{I})}$ is a σ -algebra containing the family \mathcal{I} of half-closed intervals, it necessarily contains the *smallest* σ -algebra containing the half-closed intervals. But that is to say, $\overline{R(\mathcal{I})}$ contains the Borel σ -algebra \mathcal{B} of all the Borel sets on the line.

The twin inequalities (5.4,5.5) yield more dividends. Suppose that $\{\mathbb{M}_i\}$ is a countable sequence of disjoint sets in $\overline{R(\mathcal{I})}$ and let \mathbb{M} be their union. As we've just seen \mathbb{M} is also an element of $\overline{R(\mathcal{I})}$. In view of (5.4,5.5) we have

$$F^*(\mathbb{A} \cap \mathbb{M}) + F^*(\mathbb{A} \cap \mathbb{M}^c) = \sum_{i=1}^{\infty} F^*(\mathbb{A} \cap \mathbb{M}_i) + F^*(\mathbb{A} \cap \mathbb{M}^c) \quad (5.6)$$

and, as F^* is finite, we may cancel the term $F^*(\mathbb{A} \cap \mathbb{M}^c)$ on both sides to obtain

$$F^*(\mathbb{A} \cap \mathbb{M}) = \sum_{i=1}^{\infty} F^*(\mathbb{A} \cap \mathbb{M}_i).$$

(The conclusion remains valid even if we are dealing with measures that are not finite. In this case we may not simply cancel the term $F^*(\mathbb{A} \cap \mathbb{M}^c)$ on both sides of (5.6) as it may be infinite. However, the conclusion follows again by replacing \mathbb{A} by $\mathbb{A} \cap \mathbb{M}$ in the equation as $F^*(\emptyset) = 0$ and we don't have to cancel potential infinities on both sides. This finesse is important for the case of Lebesgue measure in Section 3.) As the identity is valid for all subsets \mathbb{A} of the reals, in particular, the choice $\mathbb{A} = \mathbb{M}$ leads to the identity $F^*(\mathbb{M}) = \sum_i F^*(\mathbb{M}_i)$. It follows that, restricted to $\overline{R(\mathcal{I})}$, the set function F^* is positive and countably additive, and satisfies $F^*(\mathbb{R}) = 1$. It is natural now to focus on the restriction of F^* to the sets of the σ -algebra $\overline{R(\mathcal{I})}$.

DEFINITION 2 The set function \bar{F} is defined on the sets of $\overline{R(\mathcal{I})}$ by setting $\bar{F}(\mathbb{M}) = F^*(\mathbb{M})$ for all $\mathbb{M} \in \overline{R(\mathcal{I})}$.

Remarkably, a countably additive set function has emerged from the countably subadditive outer measure.

THEOREM 3 *The set function \bar{F} is a probability measure on the σ -algebra $\overline{R(\mathcal{I})}$ with $\bar{F}(a, b] = F(b) - F(a)$ on each half-closed interval $(a, b] \in \mathcal{I}$.*

We actually have a little more than advertised as $\overline{R(\mathcal{I})}$ is richer than the family \mathcal{B} of Borel sets. Constraining attention to the Borel sets yields the desired probability measure F .

DEFINITION 3 The set function F on the Borel sets of the line is defined uniquely as the restriction of \bar{F} (or, equivalently, F^*) to the Borel sets of the line: $F(A) = \bar{F}(A)$ for each Borel set A .

No violence is done to the earlier specification of the values of F on the ring $R(\mathcal{I})$ as the set function \bar{F} agrees with the pre-specified values of F on the ring. Our construction has hence extended the definition of F to a probability measure on the Borel sets of the line.

It only remains to show that the set function F that has been constructed so laboriously on the Borel sets is unique.

THEOREM 4 *Suppose F_1 and F_2 are probability measures on \mathcal{B} , the σ -algebra of Borel sets on the line, and suppose $F_1(a, b] = F_2(a, b] = F(b) - F(a)$ for each $a < b$. Then $F_1 = F_2$.*

PROOF: Suppose F_1 and F_2 agree on \mathcal{I} . Then they also agree on the larger family $R(\mathcal{I})$. Indeed, if $\mathbb{I} \in R(\mathcal{I})$ then there exists a finite collection of disjoint half-closed intervals $\mathbb{I}_1, \dots, \mathbb{I}_m \in \mathcal{I}$ with $\mathbb{I} = \mathbb{I}_1 \cup \dots \cup \mathbb{I}_m$. But then, by additivity,

$$F_1(\mathbb{I}) = F_1(\mathbb{I}_1) + \dots + F_1(\mathbb{I}_m) = F_2(\mathbb{I}_1) + \dots + F_2(\mathbb{I}_m) = F_2(\mathbb{I})$$

as $F_1(\mathbb{I}_j) = F_2(\mathbb{I}_j)$ for each j . Thus, the measures F_1 and F_2 agree on the ring $R(\mathcal{I})$. Now, the intersection of two half-closed intervals is either empty or another half-closed interval and so the class $R(\mathcal{I})$ of sets is closed under intersections; in the language of Section III.5, $R(\mathcal{I})$ is a π -class. As the Borel σ -algebra $\mathcal{B} = \sigma(R(\mathcal{I}))$ is generated from $R(\mathcal{I})$, this suggests an opportunity to leverage the π - λ theorem of Section III.5.

Let \mathcal{L} be the family of Borel sets A on which F_1 and F_2 agree, $F_1(A) = F_2(A)$. As we've just seen, \mathcal{L} contains the π -class $R(\mathcal{I})$. What more can be said? By the normalisation of probability measure, $F_1(\mathbb{R}) = F_2(\mathbb{R}) = 1$, and so \mathcal{L} contains \mathbb{R} as one of its elements. If $A, B \in \mathcal{L}$ and $A \subseteq B$ then, by additivity,

$$F_1(B \setminus A) = F_1(B) - F_1(A) = F_2(B) - F_2(A) = F_2(B \setminus A),$$

and so $B \setminus A$ is in \mathcal{L} . Thus, \mathcal{L} is closed under monotone unions. Finally, suppose $\{A_n, n \geq 1\}$ is an increasing sequence of elements in \mathcal{L} and $A_n \uparrow A = \bigcup_n A_n$. Then, by continuity of measure, $F_1(A_n) \rightarrow F_1(A)$ and $F_2(A_n) \rightarrow F_2(A)$. But $F_1(A_n) = F_2(A_n)$ for each n and so the limits must coincide as well, $F_1(A) = F_2(A)$. So $A \in \mathcal{L}$ and \mathcal{L} is closed under monotone unions.

In the language of Section III.5, we've established that \mathcal{L} is a λ -class containing the π -class $R(\mathcal{I})$. By the π - λ theorem it follows that $\mathcal{B} = \sigma(R(\mathcal{I})) \subseteq \mathcal{L}$ and so F_1 and F_2 agree on the Borel sets. ▶

We've hence concluded the proof of the extension theorem of Section 1: *each d.f. F induces a unique probability measure F on the Borel sets of the line.* Our construction has

actually shown a little more than advertised in that \bar{F} constitutes an extension of F to the larger σ -algebra $\bar{R}(\mathcal{I})$. See Problem 13.

A slight adaptation of the proof of Theorem 4 is needed in the case of σ -finite measures. This is required to complete the proof of Carathéodory's extension theorem for the σ -finite case, Theorem 2.3.

THEOREM 5 Suppose V is an increasing and right continuous function of a real variable. Suppose μ_1 and μ_2 are σ -finite measures on \mathcal{B} satisfying $\mu_1(a, b] = \mu_2(a, b] = V(b) - V(a)$ for each $a < b$. Then $\mu_1 = \mu_2$.

PROOF: The previous proof is impeccable if we restrict the space to a bounded half-closed interval. If $\mathbb{I} = (a, b] \in \mathcal{I}$, then μ_1 and μ_2 agree on the ring $R(\mathcal{I}) \cap \mathbb{I}$. With \mathbb{I} replacing \mathbb{R} as the space, all measures are finite, the previous proof holds verbatim, and so μ_1 and μ_2 agree on the Borel (sub) σ -algebra $\mathcal{B} \cap \mathbb{I}$. Now pick any family of bounded half-closed intervals \mathbb{I}_n which increase to \mathbb{R} ; we may pick, say, $\mathbb{I}_n = (-n, n]$. Suppose $\mathbb{A} \in \mathcal{B}$. Then $\mathbb{A} \cap \mathbb{I}_n \uparrow \mathbb{A}$ and so $\mu_1(\mathbb{A} \cap \mathbb{I}_n) \rightarrow \mu_1(\mathbb{A})$ and $\mu_2(\mathbb{A} \cap \mathbb{I}_n) \rightarrow \mu_2(\mathbb{A})$ by continuity of measure. But we've just argued that $\mu_1(\mathbb{A} \cap \mathbb{I}_n) = \mu_2(\mathbb{A} \cap \mathbb{I}_n)$ for each n and so the limits coincide, $\mu_1(\mathbb{A}) = \mu_2(\mathbb{A})$. Thus, μ_1 and μ_2 coincide on the sets of \mathcal{B} . ▶

6 Problems

1. Two-dimensional distribution functions. A continuous, positive function $F(x_1, x_2)$ on the Euclidean plane is a two-dimensional d.f. if, and only if, it increases monotonically in each argument, satisfies $F(x_1, x_2) \rightarrow 0$ if $x_1 \rightarrow -\infty$ or $x_2 \rightarrow -\infty$, $F(x_1, x_2) \rightarrow 1$ if $x_1 \rightarrow \infty$ and $x_2 \rightarrow \infty$, and the mixed difference $F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$ is positive for each choice of rectangle $(a_1, b_1] \times (a_2, b_2]$. Show that if $F(x_1, x_2)$ is a two-dimensional d.f. then $F_1(x_1) = \lim_{x_2 \rightarrow \infty} F(x_1, x_2)$ and $F_2(x_2) = \lim_{x_1 \rightarrow \infty} F(x_1, x_2)$ are one-dimensional d.f.s. The functions F_1 and F_2 are the *marginal* d.f.s associated with F .

2. Continuation. In each of the following two cases determine whether the given function $F(x_1, x_2)$ is a distribution function on the plane. If this is the case, determine the marginal distribution functions $F_1(x_1)$ and $F_2(x_2)$ of X_1 and X_2 , respectively. If this is not the case, explain why. (a) $F(x_1, x_2) = 1 - e^{-x_1-x_2}$ if $x_1, x_2 \geq 0$. (b) $F(x_1, x_2) = 1 - e^{-\min\{x_1, x_2\}} - \min\{x_1, x_2\}e^{-\min\{x_1, x_2\}}$ if $x_1, x_2 \geq 0$. The functions are to be assumed to be zero where not explicitly specified.

3. Cantor diagonalisation. In Example 3.2 we asserted that the map $C: (m, n) \mapsto m + (m+n)(m+n+1)/2$ from pairs of positive integers into the natural numbers is one-to-one. Prove this.

4. Countability. If A and B are countable sets, the Cartesian product $A \times B$ consisting of ordered pairs (a, b) with $a \in A$ and $b \in B$ is countable. If A_1, \dots, A_n are countable sets then the Cartesian product $A_1 \times \dots \times A_n$ is countable. The set of sequences (a_1, a_2, \dots) where, for each i , a_i is an element of a denumerably infinite set A_i , is not countable.

5. More on countability. Use a diagonalisation argument to show that the union of a countable number of disjoint countable sets is countable. Hence show that the union of a countable number of countable sets is countable.



6. *Lebesgue-measurable sets that are not Borel.* For the reader familiar with cardinal arithmetic: as the family \mathcal{B} of Borel sets may be generated by the family of open intervals with rational centres, argue that \mathcal{B} has the cardinality c of the continuum. As the Cantor set \mathbb{C} of Example 3.3 has Lebesgue measure zero all its subsets have Lebesgue measure zero. As $2^c > c$ argue that most subsets of \mathbb{C} are Lebesgue-measurable but not Borel.

7. *Translation invariance of Lebesgue measure.* For any subset \mathbb{A} of the real line and any real r define the set $\mathbb{A} + r$, the *translate of \mathbb{A} by r* , to be the collection of points $x + r$ where $x \in \mathbb{A}$. Show that if \mathbb{A} is Lebesgue-measurable then so too is $\mathbb{A} + r$ and $\lambda(\mathbb{A} + r) = \lambda(\mathbb{A})$.

8. *A non-measurable set.* If x and y are real, write $x \sim y$ if $x - y = q$ is rational. The relation \sim determines, for each x , an *equivalence class* of points y with $x \sim y$. Select a point $x \in (0, 1)$ in each distinct equivalence class, the collection of these points forming a set \mathbb{E} . (For the reader who knows it, the assertion that we can form such a set is a consequence of the axiom of choice.) Show that if $x \in (0, 1)$ then $x \in \mathbb{E} + q$ for some rational $q \in (-1, 1)$.

9. *Continuation.* Show that if q and r are distinct rationals then the sets $\mathbb{E} + q$ and $\mathbb{E} + r$ are disjoint.

10. *Continuation.* Let $\mathbb{S} = \bigcup_q (\mathbb{E} + q)$, the union ranging over all rationals q in $(-1, 1)$. Assuming that \mathbb{S} is Lebesgue-measurable show hence that $\lambda(\mathbb{S})$ is either 0 or ∞ . On the other hand, argue that $(0, 1) \subseteq \mathbb{S} \subseteq (-1, 2)$ and hence that $1 \leq \lambda(\mathbb{S}) \leq 3$. Conclude that \mathbb{S} is not measurable.

11. *Approximation.* Let F be any d.f., \mathbb{I} any Borel set, and ϵ any strictly positive quantity. Using the definition of outer measure show that there exists a finite union of half-closed intervals, $\mathbb{I}_0 = \bigcup_{i=1}^n (a_i, b_i]$, such that $F(\mathbb{I} \Delta \mathbb{I}_0) \leq \epsilon$. Show that the conclusion holds also for Lebesgue measure λ if $\lambda(\mathbb{I}) < \infty$.

12. *Continuation.* Let \mathbb{A} be any Borel set. Then there exists a Borel set $\mathbb{B} \supseteq \mathbb{A}$ with $F(\mathbb{A}) = F(\mathbb{B})$ where $\mathbb{B} = \bigcap_n \mathbb{B}_n$ is the limit of a decreasing family of sets $\mathbb{B}_1 \supseteq \mathbb{B}_2 \supseteq \dots \supseteq \mathbb{B}_n \supseteq \dots$ each of which may be expressed as a countable union of disjoint half-closed intervals. Moreover, for each n , we may express $\mathbb{B}_n = \bigcup_k \mathbb{B}_{nk}$ as the limit of an increasing family of sets $\mathbb{B}_{n1} \subseteq \mathbb{B}_{n2} \subseteq \dots \subseteq \mathbb{B}_{nk} \subseteq \dots$ each of which is a finite union of disjoint half-closed intervals.



13. *Completion.* Let F be a d.f. and F the induced probability measure on the σ -algebra \mathcal{B} of Borel sets of the line. Let $\bar{\mathcal{B}}$ be the completion of \mathcal{B} with respect to F and let \bar{F} be the completion of the measure. Show that \bar{F} is identical to the outer measure F^* on $\overline{R(J)}$. Thus, completing the measure F results in the same measure \bar{F} obtained in Section 5.

XII

Random Variables

Chance experiments on the continuum lead naturally to outcomes that one may call random variables. From a more general vantage point we may consider any quantitative measurement of some attribute of the sample points of an abstract experiment to represent induced chance variables.

1 Measurable maps

We turn to an abstract sample space Ω representing the idealised outcomes of a chance experiment. We equip Ω with a σ -algebra \mathcal{F} of subsets of Ω constituting the family of events of interest. The pair (Ω, \mathcal{F}) is called a *measurable space* though the term is a misnomer as no measure has yet been specified. A real-valued function $X(\omega)$ mapping the abstract sample points $\omega \in \Omega$ into the real line represents an attempt to quantify or make measurable certain features of the underlying chance experiment. The reader is familiar with the case when the sample space is the real line and we deal with functions of the coordinate variable. The following examples illustrate the value of a more abstract point of view.

EXAMPLES: 1) *Three tosses of a coin.* The sample points of the experiment may be identified with triples $\omega = (\omega_1, \omega_2, \omega_3)$ where $\omega_j \in \{\text{H}, \text{T}\}$ connotes the outcome of the j th toss. If we associate with each sample point ω the number of successes we obtain the map X tabulated below.

ω	HHH	HHT	HTH	HTT	THH	HTT	TTH	TTT
$X(\omega)$	3	2	2	1	2	1	1	0

The natural probability distribution on the sample space is uniform in the eight possibilities whence $X(\omega)$ inherits the binomial distribution $b_3(\cdot; 1/2)$ corresponding to accumulated successes in three tosses of a fair coin.

2) *Unending sequence of coin tosses.* We have seen the sample space corresponding to an infinitely prolonged sequence of tosses in Example I.7.7. In this case,

as we have seen, we may identify each sample point $\omega = (\omega_k, k \geq 1)$ (where each $\omega_k \in \{\mathfrak{H}, \mathfrak{T}\}$) with a point $t = \sum_k z_k 2^{-k}$ in the unit interval $[0, 1]$ (where $z_k = 1$ if $\omega_k = \mathfrak{H}$ and $z_k = 0$ if $\omega_k = \mathfrak{T}$). The map X which assigns to each ω the value $\lim_n n^{-1}(z_1 + \dots + z_n)$ if the limit exists, and the value -1 otherwise, maps each sample point into the density of 1s in a dyadic expansion if the density exists. As ω ranges over the entire sample space of all possible sequences of coin tosses, the values $X(\omega)$ range over the entire continuum of values in the unit interval with the point $\{-1\}$ appended. Nevertheless, by Borel's law of normal numbers (Section V.7), the distribution inherited by X is degenerate and places all its mass at the point $1/2$. In consequence, the probability is identically zero that $X(\omega)$ takes a value in any interval not containing the point $1/2$; the sample point corresponding to an infinite repetition of $\mathfrak{H}\mathfrak{H}\mathfrak{T}\mathfrak{H}\mathfrak{H}\mathfrak{T}\dots$ is mapped by X into the value $2/3$ but the probability of occurrence of this point is zero.

3) *The random graph $G_{n,p}$.* A graph $G = (V, E)$ consists of a collection of vertices $V = \{1, \dots, n\}$ together with a subcollection of unordered pairs of vertices $\{i, j\}$ called edges. We say that two vertices i and j are *adjacent* if $\{i, j\} \in E$. A *colouring* of G assigns to each vertex i a colour such that each pair of adjacent vertices i and j have different colours. The *chromatic number* $\chi(G)$ is the smallest number of colours needed to colour the graph G .

Colouring a map of the United States so that no two adjacent states have the same colour and allocating frequency bands in a wireless communications system so that adjacent users or cells do not use the same frequencies are both examples of graph colourings. In these and a ubiquitous range of applications the chromatic number of a graph captures parsimony in the allocation of resources.

For each n , the chromatic number χ is a map from the family of graphs on n vertices into the numbers $\{1, \dots, n\}$. In this setting the Erdős–Rényi random graph $G_{n,p}$ examined in Example III.4.5 provides a natural chance mechanism for the random selection of a graph. The associated chromatic number $\chi(G_{n,p})$ has a distribution with support in the integers $\{1, 2, \dots, n\}$ though it is not at all trivial to characterise it.

A variety of other natural functions that are informative about graph structure may be formulated. Among them: the *independence number* of a graph is the size of the largest set of vertices no two of which are adjacent. The *degree* of a vertex is the number of vertices adjacent to it; the *minimum degree* of a graph is the smallest degree of any of its vertices. A *cycle* is a sequence of vertices $(u = i_1, i_2, \dots, i_{k-1}, i_k = u)$ where the first and last vertices are the same, no other vertices are repeated, and such that $\{i_j, i_{j+1}\}$ is an edge for each j ; the *girth* of a graph is the length of its longest cycle.

4) *The bin-packing problem.* Suppose U_1, \dots, U_n represent flows of commodities, say, traffic from different sources in the internet where we suppose that the

granularity of specification is so fine that we may treat traffic as a continuum flow. Assuming bounded flows we may suppose by a suitable normalisation that $0 \leq U_k \leq 1$ for each k . Given a set of unit-capacity bins the resource allocation problem deals with assigning flows to bins. Multiple flows may be assigned to a given bin but with the proviso that flows may not be split across bins. In this context the *binning number* $X(U_1, \dots, U_n)$ defined to be the smallest number of bins which can accommodate the given flows provides a benchmark for how efficiently resources are utilised.

The unit cube $[0, 1]^n$ in n dimensions is the natural sample space in this setting, each sample point $\mathbf{U} = (U_1, \dots, U_n)$ representing a vector of flows. If we assume that distinct flows are independent and uniformly distributed then the uniform density in the cube confers the appropriate measure on the Borel sets of the cube, the probability that \mathbf{U} takes a value in any Borel set A given by $\text{Vol}(A) = \int_A d\mathbf{u}$. (There is no difficulty in interpreting the integral as volume if A is a rectangle in $[0, 1]^n$; for a general Borel set the reader should temporarily interpret the expression as simply a vivid representation of the Lebesgue measure of the set, $\lambda(A)$, which, by Carathéodory's theorem, is the unique extension of volume to general Borel sets.) The binning number $X: [0, 1]^n \rightarrow \{1, \dots, n\}$ then constitutes a fixed function of the random vector of flows.

5) *Finite-energy functions.* To each square-integrable function f we may associate its energy $X(f) = \int |f(x)|^2 dx$. The map X is a real-valued function defined on the space L^2 of square-integrable functions. Functions $f \in L^2$ are the sample points, their energy $X(f)$ the measurable quantity of interest. While the observable energy is of natural interest in this setting, the formulation of a measurable function space—and the specification of probability measure on this space—is a matter of some subtlety. ▶

We begin accordingly with some abstract measurable space (Ω, \mathcal{F}) that we consider to have been specified (if only implicitly) and consider the nature of a real-valued function $X(\omega)$ defined on the space. With what properties should the map $\omega \mapsto X(\omega)$ be equipped? Well, we will almost definitely wish to associate probabilities with intervals on the line so that intervals provide a natural starting point. As before, we start with the family \mathcal{I} of bounded, half-closed intervals of the form $(a, b]$.

For any subset \mathbb{I} of the real line, the *preimage under X of \mathbb{I}* , denoted $X^{-1}\mathbb{I}$, represents the collection of points ω for which $X(\omega) \in \mathbb{I}$. In particular, if $\mathbb{I} = (a, b]$ is a half-closed interval then $X^{-1}(a, b] = \{\omega : a < X(\omega) \leq b\}$.

DEFINITION A real-valued function X on Ω is *measurable* (with respect to the σ -algebra \mathcal{F}) if $X^{-1}(a, b] \in \mathcal{F}$ for every half-closed interval $(a, b]$ in \mathcal{I} .

When it is necessary to identify the measurable space (Ω, \mathcal{F}) in question we say that X is *\mathcal{F} -measurable*. If the underlying sample space is the real line and the associated σ -algebra is the family of Borel sets on the line then we say simply that

X is *Borel-measurable*. It will be convenient on occasion to use this terminology also for extended real-valued functions that map Ω into the real line expanded by addition of the points $-\infty$ and $+\infty$.

If the σ -algebra \mathcal{F} is equipped with a probability measure P then X carries over probabilities to intervals and thence, as we shall see, to Borel sets on the line. In a more vivid language, a real-valued measurable map X on a probability space is called a *random variable*. This is a hideous misuse of language though the terminology, regrettably, is now entrenched. Its origins date to a simpler time when the outcome of the experiment could be considered to be a real number X governed by an underlying probability law; that is to say the experimental outcome X was a random variable in the sense that we would associate in ordinary language.¹ In the general setting where Ω is an abstract sample space, however, $X: \Omega \rightarrow \mathbb{R}$ represents a measurable map. Viewed as a real-valued map on the sample space there is nothing random at all about X (though in practice it may only be implicitly specified); it is the sample point ω that is the argument of X that is randomly selected and in consequence it is the *value* $X(\omega)$ that X takes at the point ω that is *a priori* unknown and governed by a probability law inherited from the underlying probability measure. It is now, naturally enough, the probability law governing the values $X(\omega)$ that is of immediate interest.

To each interval $(a, b]$ in \mathcal{I} is associated the event $X^{-1}(a, b] = \{\omega : a < X(\omega) \leq b\}$ in the σ -algebra \mathcal{F} . Accordingly, to the interval $(a, b]$ we may associate the probability $P(X^{-1}(a, b]) = P\{\omega : a < X(\omega) \leq b\}$. How about other types of intervals and, more generally, Borel sets \mathbb{I} on the line? We first have to negotiate the difficulty that it is not immediately clear whether the preimage under X of a Borel set \mathbb{I} is in general an element of the σ -algebra \mathcal{F} . As we may assign probabilities only to events, that is to say, elements of \mathcal{F} , it is clear that we have to consider this issue first.

MEASURABILITY THEOREM *Suppose X is measurable with respect to a σ -algebra \mathcal{F} . Then $X^{-1}\mathbb{I} \in \mathcal{F}$ for every Borel set \mathbb{I} on the line.*

PROOF: The demonstration rests upon the fact that the Borel σ -algebra \mathcal{B} is the *smallest* σ -algebra containing the family of half-closed intervals \mathcal{I} .

Let \mathcal{G} be the family of all subsets \mathbb{I} of the real line for which $X^{-1}\mathbb{I} \in \mathcal{F}$. By definition the family of half-closed intervals \mathcal{I} is contained in \mathcal{G} as X is measurable with respect to \mathcal{F} . It will suffice now to show that \mathcal{G} is itself a σ -algebra of subsets of the real line \mathbb{R} , that is, that \mathcal{G} contains \mathbb{R} and is closed under complementation and countable unions. As the family \mathcal{B} of Borel sets

¹In Chapter VII we dealt with probability experiments on the continuum $\Omega = \mathbb{R}$ and X served in the rôle of a coordinate variable. From our new vantage point we may think of the coordinate random variable as the identity map $X: \omega \mapsto \omega$ on a real-line sample space.

is the *smallest* σ -algebra containing \mathcal{I} it must then follow that \mathcal{G} contains \mathcal{B} and hence that the preimage of every Borel set is an event.

As $X^{-1}\mathbb{R} = \Omega$ and the sample space Ω is an element of \mathcal{F} , it follows that the whole line \mathbb{R} is in \mathcal{G} .

Suppose \mathbb{I} is a subset of the line contained in the family \mathcal{G} . Then $X^{-1}\mathbb{I} \in \mathcal{F}$ whence $(X^{-1}\mathbb{I})^c \in \mathcal{F}$ as \mathcal{F} is closed under complementation. But the complement of $X^{-1}\mathbb{I}$ is the set of precisely those sample points ω which are not mapped into \mathbb{I} by X . Accordingly, $(X^{-1}\mathbb{I})^c = X^{-1}(\mathbb{I}^c)$ and $\mathbb{I}^c \in \mathcal{G}$. It follows that \mathcal{G} is closed under complementation.

Finally, suppose $\{\mathbb{I}_n\}$ is a countable sequence of subsets of the line in \mathcal{G} . Then $X^{-1}\mathbb{I}_n \in \mathcal{F}$ for each n whence $\bigcup_n X^{-1}\mathbb{I}_n \in \mathcal{F}$ as \mathcal{F} is closed under countable unions. But the union of the sets $X^{-1}\mathbb{I}_n$ is the set consisting of those sample points ω which are mapped into at least one of the sets \mathbb{I}_n . In other words, $\bigcup_n X^{-1}\mathbb{I}_n = X^{-1}(\bigcup_n \mathbb{I}_n)$ whence $\bigcup_n \mathbb{I}_n$ is in \mathcal{G} . ▶

Write $\sigma(X)$ for the family of events that may be expressed in the form $X^{-1}\mathbb{A}$ as \mathbb{A} ranges over all Borel sets of the line. The proof of our measurability theorem shows then that $\sigma(X)$ is indeed a σ -algebra of events. The family $\sigma(X)$ is manifestly contained in \mathcal{F} and, in general, is smaller than \mathcal{F} . We call $\sigma(X)$ the σ -algebra *generated by* X .

It is common to use the conclusion of our measurability theorem as the *definition* of a measurable function. But as a general principle this author has found that it is preferable to keep a definition as simple as possible rather than endow it with a patina of spurious generality. From this perspective it has seemed worthwhile to cast the definition of measurability in terms of intervals rather than in terms of general Borel sets for which we have far less intuition. Our measurability theorem tells us that no flexibility is lost with this approach.

2 The induced measure

Suppose that X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. In view of the measurability theorem of the previous section, to each Borel set \mathbb{I} we may in principle associate the probability $\mathbf{P}(X^{-1}\mathbb{I}) = \mathbf{P}\{\omega : X(\omega) \in \mathbb{I}\}$ (though there is still the small matter of how to do the computation). The notation is unnecessarily pedantic and there is no risk of confusion if we suppress the underlying sample point ω in the notation and write simply $\mathbf{P}\{X \in \mathbb{I}\}$ for the probability that $X(\omega)$ takes a value in the set \mathbb{I} .

As the Borel sets contain all the intervals, finite as well as infinite, it follows in particular that we may associate probabilities to intervals of any type and the expressions $\mathbf{P}\{a < X < b\}$, $\mathbf{P}\{a \leq X \leq b\}$, $\mathbf{P}\{a \leq X < b\}$, and $\mathbf{P}\{a < X \leq b\}$ are all perfectly legitimate and stand for the probabilities that $X(\omega)$ takes a value in the intervals (a, b) , $[a, b]$, $[a, b)$, and $(a, b]$, respectively. Their infinite counterparts are expressions like $\mathbf{P}\{X < a\}$, $\mathbf{P}\{X \leq a\}$, $\mathbf{P}\{X \geq b\}$, and $\mathbf{P}\{X > b\}$

representing the probabilities that $X(\omega)$ takes a value in the infinite intervals $(-\infty, a]$, $(-\infty, a]$, $[b, \infty)$, and (b, ∞) , respectively.

While we may thus, in principle, refer the computation of the probability that X takes value in some Borel set \mathbb{I} back to the underlying probability space $(\Omega, \mathcal{F}, \mathbf{P})$, in practice this is an unwieldy procedure and we look naturally enough for a direct characterisation of the probability measure induced on the Borel sets of the line by X .

The probability that $X(\omega)$ takes value in an interval $(-\infty, x]$ depends only on the right endpoint x and hence determines a real-valued function $F(x)$ of a real variable x via the correspondence $F: x \mapsto \mathbf{P}\{X \leq x\}$. Laid out this says

$$F(x) = \mathbf{P}\{X \leq x\} = \mathbf{P}\{\omega : X(\omega) \leq x\} = \mathbf{P}(X^{-1}(-\infty, x]),$$

the differing expressions all being just notational variants of each other.

THEOREM *The function F is right continuous, increases monotonically, and has limits $F(-\infty) = 0$ and $F(\infty) = 1$. In other words, F is a distribution function.*

PROOF: Suppose $\{x_n\}$ is any sequence of real values decreasing to a limiting value x , in notation, $x_n \downarrow x$. Then $\{X^{-1}(-\infty, x_n]\}$ forms a nested sequence of sets decreasing to the limiting set $\bigcap_n X^{-1}(-\infty, x_n] = X^{-1}(\bigcap_n (-\infty, x_n]) = X^{-1}(-\infty, x]$, whence $F(x_n) = \mathbf{P}(X^{-1}(-\infty, x_n]) \rightarrow \mathbf{P}(X^{-1}(-\infty, x]) = F(x)$ by continuity of probability measure. Thus F is right continuous.

Now suppose $x \leq y$. Then the set $\{\omega : X(\omega) \leq x\}$ is contained in the set $\{\omega : X(\omega) \leq y\}$ whence $F(x) = \mathbf{P}\{\omega : X(\omega) \leq x\} \leq \mathbf{P}\{\omega : X(\omega) \leq y\} = F(y)$ by monotonicity of probability measure. It follows that F increases monotonically.

Finally, to establish the limits at $\pm\infty$, suppose first that $\{x_n\}$ is a sequence decreasing to $-\infty$, in notation, $x_n \downarrow -\infty$. Then $\bigcap_n X^{-1}(-\infty, x_n] = X^{-1}(\bigcap_n (-\infty, x_n]) = X^{-1}\emptyset = \emptyset$ and another appeal to continuity of probability measure shows that $F(x_n) = \mathbf{P}(X^{-1}(-\infty, x_n]) \rightarrow \mathbf{P}(\emptyset) = 0$. Likewise, if $\{x_n\}$ is any unbounded increasing sequence, $x_n \uparrow \infty$, then the corresponding sequence $\{X^{-1}(-\infty, x_n]\}$ forms another nested sequence of sets, increasing this time, with limit $\bigcup_n X^{-1}(-\infty, x_n] = X^{-1}(\bigcup_n (-\infty, x_n]) = X^{-1}\mathbb{R} = \Omega$ so that $F(x_n) = \mathbf{P}(X^{-1}(-\infty, x_n]) \rightarrow \mathbf{P}(\Omega) = 1$ by a final appeal to continuity. ▶

The value of a consideration of the distribution function F of the random variable X becomes clear when we make the observation that, by additivity of probability measure, $\mathbf{P}\{a < X \leq b\} = \mathbf{P}\{X \leq b\} - \mathbf{P}\{X \leq a\} = F(b) - F(a)$, so that the d.f. F determines a positive set function \mathbf{F} on the family \mathcal{J} of half-closed intervals defined by $\mathbf{F}(a, b] = F(b) - F(a)$ for each $a \leq b$; we identify this set function with the probabilities (inherited from the underlying probability measure) of the various half-closed intervals. But, as asserted in the extension theorem of Section XI.1, there is a unique probability measure \mathbf{F} on the Borel sets of the line that is consistent with the map $\mathbf{F}: (a, b] \mapsto F(b) - F(a)$ on the

half-closed intervals. And thus, all probabilities relevant to the random variable X may be written directly in terms of this induced measure: $P\{X \in \mathbb{I}\} = F(\mathbb{I})$ for every Borel set \mathbb{I} on the line. In particular, $P\{a < X \leq b\} = F(a, b]$, $P\{a < X < b\} = F(a, b)$, $P\{a \leq X < b\} = F[a, b)$, and $P\{a \leq X \leq b\} = F[a, b]$ where we have to make a distinction between the types of intervals to allow for situations where X puts mass at an interval endpoint. The infinite counterparts of these relations then become $P\{X \leq b\} = F(-\infty, b]$, $P\{X < b\} = F(-\infty, b)$, $P\{X > a\} = F(a, \infty)$, and $P\{X \geq a\} = F[a, \infty)$. The right continuity of the d.f. allows us to write probabilities in these simple cases directly in terms of the d.f. Thus, $F(a, b] = F(b) - F(a)$, $F(a, b) = F(b-) - F(a)$, $F[a, b] = F(b) - F(a-)$, and $F[a, b) = F(b-) - F(a-)$, the limiting values of the d.f. taking care of the situations when a tends to $-\infty$ or b tends to $+\infty$.

The d.f. F hence determines probabilities for all types of intervals and, in general, all Borel sets through the probability measure F associated with it. Of course, any given probability measure F on the Borel sets also uniquely determines a corresponding d.f. via the correspondence $F(x) = F(-\infty, x]$. Thus, the probabilities on the line are determined by the d.f. and vice versa. *In consequence the typographic distinction between the d.f. F and the associated probability measure F need not be maintained and we may consolidate and simplify notation by using the common notation F for both entities.* The argument of F will determine whether we intend its use as an ordinary function of a real variable as in the d.f. $F(x)$ or whether it appears in the rôle of a set function on the Borel sets as in the probability measure $F(\mathbb{I})$. There is no great risk of confusion in abusing notation in this fashion—the context determines in what sense we are referring to the function—; and on the positive side, the consolidation of notation both simplifies the typography and reinforces the idea that the d.f. (at least implicitly) specifies the induced probability measure.

Thus, to each random variable X we may associate a d.f. (probability measure) F , and conversely, to each d.f. (probability measure) F we may associate a random variable X (the coordinate variable for the probability space $(\mathbb{R}, \mathcal{B}, F)$ will serve). We say that F (in either rôle) is the *distribution* or *law* of X .

Once the d.f. F of a random variable X is specified, the rôle of the originating probability space may be obscured without loss as all probabilities relevant to X may be determined from F . From this point on we may deal with the random variable X as if it were a coordinate variable on a real-line sample space governed by the distribution F . This suggests that we examine distribution functions in more detail.

3 Discrete distributions

Suppose F is any d.f. As F is bounded and increasing the limits $F(x-)$ and $F(x+)$ both exist for each x ; and as F is right continuous $F(x+) = F(x)$ for each x . If F

is continuous at the point x then the left and right limits at x will coincide and we have $F(x-) = F(x)$. If, on the other hand, $F(x-) < F(x)$ then F has a *jump discontinuity* at the point x and the positive value $F(x) - F(x-)$ is called the *size of the jump at x* . As F is increasing all its points of discontinuity, if any, must be jumps and these then are of interest. (The reader should mull over what kinds of discontinuities there are in general.)

Let \mathbb{J} denote the points of jump x of the d.f. F . The next result shows by a useful topological argument that the set \mathbb{J} cannot be overly large.

THEOREM *The set of discontinuities of a d.f. F is countable.*

PROOF: To each point of jump x of F we may associate the open interval $\mathbb{I}_x = (F(x-), F(x))$. We claim that $\{\mathbb{I}_x, x \in \mathbb{J}\}$ is a disjoint family of open intervals. Indeed, suppose x and y are two points of jump of F and suppose $x < y$. By monotonicity of F we then have $F(x-) < F(x) \leq F(y-) < F(y)$ so that \mathbb{I}_x and \mathbb{I}_y are non-intersecting. It follows that the family $\{\mathbb{I}_x, x \in \mathbb{J}\}$ is disjoint. As each open interval of the real line contains rational points, to each \mathbb{I}_x we may associate a rational number q_x contained in it, none of these rational numbers coinciding as the intervals $\{\mathbb{I}_x\}$ are disjoint. Thus to each point of jump we may associate a distinct rational number; equivalently, we have constructed a one-to-one map from the set of discontinuities of F into the set of rational numbers \mathbb{Q} . But the set of rational numbers is countable so that \mathbb{J} is as well. ▶

If the reader will consult Problem 1 she will find an alternative proof via a counting argument of equal utility.

Thus, if \mathbb{J} is non-empty we may enumerate it (in some order) as a countable collection $\{x_j\}$ of points of jump of F . This allows us to make an initial characterisation. Let $H_0(x)$ denote the *Heaviside* or *unit step* function which takes value 1 if $x \geq 0$ and value 0 if $x < 0$.

DEFINITION A *discrete d.f.* F is one that may be represented in the form

$$F(x) = \sum_j p_j H_0(x - x_j) \quad (3.1)$$

where $\{x_j\}$ is any countable collection of real numbers, p_j is positive for each j , and $\sum_j p_j = 1$. We refer to the sequence $\{p_j\}$ as the associated *distribution* and say that a random variable is discrete if its distribution function is discrete.

We should first verify that a function of the form (3.1) does represent a bona fide d.f. As $H_0(x - x_j)$ is equal to 1 if $x \geq x_j$ and is equal to 0 otherwise, for any x , the sum picks out just those terms p_j for which $x_j \leq x$. For any function F of this form it is hence clear that $F(-\infty) = 0$ and $F(\infty) = \sum_j p_j = 1$. Furthermore, F is increasing as each of the functions $H_0(x - x_j)$ is. To verify that F is a d.f. it hence suffices to establish that it is right continuous. But this follows

immediately because each of the functions $H_0(x - x_j)$ is right continuous and the series on the right of (3.1) converges *uniformly*.

For the reader who is not convinced by this airy assertion, in detail, the uniform convergence of the series means that, for every $\epsilon > 0$, there exists n such that

$$\left| F(x) - \sum_{|j| \leq n} p_j H_0(x - x_j) \right| = \left| \sum_{|j| > n} p_j H_0(x - x_j) \right| \leq \sum_{|j| > n} p_j < \epsilon$$

uniformly for all x . (This follows because of the convergence of the series $\sum_j p_j$.) But the right continuity of the Heaviside function implies that we can find $t_0 > x$ such that

$$\left| \sum_{|j| \leq n} p_j H_0(t - x_j) - \sum_{|j| \leq n} p_j H_0(x - x_j) \right| = \sum_{|j| \leq n} p_j |H_0(t - x_j) - H_0(x - x_j)| < \epsilon$$

for all $x < t < t_0$. Two applications of the triangle inequality then show that

$$\begin{aligned} |F(t) - F(x)| &\leq \left| F(t) - \sum_{|j| \leq n} p_j H_0(t - x_j) \right| + \left| \sum_{|j| \leq n} p_j H_0(t - x_j) - \sum_{|j| \leq n} p_j H_0(x - x_j) \right| \\ &\quad + \left| \sum_{|j| \leq n} p_j H_0(x - x_j) - F(x) \right| < 3\epsilon. \end{aligned}$$

As the positive ϵ may be chosen arbitrarily small, it follows that $F(t) \rightarrow F(x)$ as $t \downarrow x$, as was to be shown.

In the representation (3.1), the collection $\{x_j\}$ enumerates all the points of jump of F with each $p_j = F(x_j) - F(x_j^-)$ representing the size of the jump of F at the point x_j . To verify this we observe that

$$F(x) - F(x^-) = \sum_j p_j [H_0(x - x_j) - H_0(x^- - x_j)]$$

again as the series defining F is uniformly convergent. If $x \neq x_j$ for any j then each of the summands vanishes and the series converges to 0; if, on the other hand, $x = x_k$ for some k , then the only term on the right that does not vanish is the term corresponding to $j = k$ and the series on the right converges to the value p_k . Thus, F has jumps only at the points x_j and at each such point the size of the jump is p_j . Thus a discrete d.f. is characterised by the property that its points of jump are the only points of increase of F . If X is a random variable associated with the discrete d.f. F of (3.1) then X takes values only in the countable set $\{x_j\}$ with $P\{X = x_j\} = F(x_j) - F(x_j^-) = p_j$ for each j . If the collection $\{x_j\}$ is a subset of the integers then X is *arithmetic*. In a suggestive language, the points of jump of a discrete d.f. are also called its *atoms*.

EXAMPLES: 1) *Degenerate d.f.* The function $F(x) = H_0(x)$ itself determines a distribution which places all its mass at the origin; this is the *Heaviside distribution*. More generally, we say that a discrete d.f. F is *degenerate* if it is of the form $F(x) = H_0(x - x_0)$ for some x_0 . A random variable with this distribution takes only the fixed value x_0 with probability one.

2) *Indicator random variables.* Suppose A is any event in an abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The *indicator of A* is the random variable 1_A which takes value 1 if A occurs and value 0 if A does not occur; or, explicitly, $1_A(\omega) = 0$ if $\omega \notin A$ and $1_A(\omega) = 1$ if $\omega \in A$. The value taken by $1_A(\omega)$ indicates whether or not A has occurred, hence the name. The indicator notation, while a little startling at first blush, at least has the virtue of being vivid. We will use it in other settings as well.

An indicator random variable represents a metaphorical coin toss which turns up heads if, and only if, the event A occurs. Thus, the *indicator variable* 1_A represents a Bernoulli trial with success probability $p = \mathbf{P}(A)$ and failure probability $q = 1 - \mathbf{P}(A)$. Its d.f. $F(x) = qH_0(x) + pH_0(x - 1)$ has a jump of size q at $x = 0$ and a jump of size p at $x = 1$.

3) *Simple random variables.* A random variable is *simple* if it can be represented as a finite superposition of indicator variables. In detail, X is simple if it has a representation of the form $X(\omega) = x_1 1_{A_1}(\omega) + \dots + x_n 1_{A_n}(\omega)$ where A_1, \dots, A_n is a finite collection of events and x_1, \dots, x_n are arbitrary real numbers. If we write $p_i = \mathbf{P}(A_i)$ for each i then X takes value x_i with probability p_i and the d.f. has the form $F(x) = p_1 H_0(x - x_1) + \dots + p_n H_0(x - x_n)$.

Any discrete random variable taking only a finite number of values is simple. Suppose X takes only values x_1, \dots, x_n with positive probability. For each j , we may now identify $A_j = X^{-1}\{x_j\}$, that is, A_j is the preimage under X of the set consisting of the singleton x_j .

4) *The binomial distribution.* The number of successes in n tosses of a coin with success probability p is simple with d.f. $F(x) = \sum_{k=0}^n b_n(k; p) H_0(x - k)$ where $b_n(k; p) = \binom{n}{k} p^k (1 - p)^{n-k}$ represents the binomial probabilities.

5) *The Poisson distribution.* If X is Poisson with mean α then its d.f. has atoms at integer $k \geq 0$ and is given by $F(x) = \sum_{k=0}^{\infty} p(k; \alpha) H_0(x - k)$. The sizes of the jumps $p(k; \alpha) = e^{-\alpha} \alpha^k / k!$ are the Poisson probabilities.

6) *Rational atoms.* Let x_1, x_2, \dots be any enumeration of the rational numbers and let p_1, p_2, \dots be any sequence of positive numbers adding to one. (For definiteness, we may take $p_j = 2^{-j}$ for $j \geq 1$.) The corresponding d.f. has the form (3.1) with jumps only at the rational points x_j . The associated random variable takes only rational values with positive probability. This example shows that *the set of points of jump of a discrete d.f. may be dense in the real line*, that is to say, every interval of the real line can contain points of jump. ▶

4 Continuous distributions

A d.f. F is *continuous* if it has no jumps. This is just the ordinary notion of continuity of a function: a continuous d.f. is a d.f. that is continuous everywhere

on the line. The simplest case is when the d.f. F is sufficiently smooth to enjoy a positive derivative.

DEFINITION 1 A d.f. F is *absolutely continuous* if it can be represented in the form

$$F(x) = \int_{-\infty}^x f(t) dt \quad (4.1)$$

where f is a positive, integrable function with $\int_{-\infty}^{\infty} f(t) dt = 1$ called the *density* of F (or of the associated random variable X). We also say that a random variable X is absolutely continuous if its d.f. is.

The reader should note that we may express (4.1) in the form

$$F(x) = \int_{-\infty}^{\infty} f(t)H_0(x-t) dt, \quad (4.1')$$

the upper limit of integration only formally infinite as $H_0(x-t)$ is identically zero for $t > x$. In this form it becomes apparent that (4.1') is just the continuous analogue of (3.1). As f is assumed integrable, a variation of the previous argument shows that a function of the form (4.1') is indeed a d.f.

For the time being we can take the term “integrable” at face value in the Riemann sense though we will provide a much more flexible and powerful interpretation in the next chapter. Certainly there is no problem in the interpretation if f is, say, piecewise continuous, in which case (4.1) provides an unambiguous specification of the d.f. and hence the associated probability measure. While there is no difficulty in the interpretation of an expression like $\int_a^b f(t) dt$ as the area under the curve of f in the region $a < t < b$, an expression like $\int_{\mathbb{A}} f(t) dt$ is harder to interpret within the confines of the Riemann theory of integration when \mathbb{A} is a general Borel set. For the nonce, given a density f , we use (4.1) as the defining relation of the associated absolutely continuous d.f. F , and, when \mathbb{A} is a general Borel set, we agree to interpret $\int_{\mathbb{A}} f(t) dt$ as simply $F(\mathbb{A})$, the probability measure attached to \mathbb{A} by the d.f.

The setting is certainly familiar even if the necessity for the caveat “absolutely” in the definition is not immediately apparent.

EXAMPLES: 1) *The uniform distribution.* The uniform density $u(x) = 1$ for $0 < x < 1$ has support in the unit interval $(0, 1)$ only. The corresponding d.f. increases linearly in the unit interval, $U(x) = x$ for $0 < x < 1$. (And, of course, $U(x) = 0$ for $x \leq 0$ and $U(x) = 1$ for $x \geq 1$.) The probability measure induced by this d.f. is Lebesgue measure on the unit interval that we had encountered in Section V.3 and again in Section XI.3.

2) *The exponential distribution.* The exponential density with mean $1/\alpha$ has support on the positive half-line and has the form $g(x) = \alpha e^{-\alpha x}$ for $x \geq 0$. The corresponding d.f. is given by $G(x) = 1 - e^{-\alpha x}$ for $x \geq 0$.

3) *The normal distribution.* The standard normal density $\phi(x) = e^{-x^2/2}$ has corresponding d.f. $\Phi(x) = \int_{-\infty}^x \phi(t) dt$. As we have seen, the expression for the d.f. cannot be further simplified in terms of other elementary functions. ▶

A given density f uniquely determines the associated absolutely continuous d.f. F via the defining relation (4.1). On the flip side, the first fundamental theorem of calculus tells us that F is differentiable, at least at points of continuity of f , and at these points $F'(x) = f(x)$. Thus, the associated density is effectively determined by the d.f. as well. This still leaves open the question, however, of whether every continuous d.f. is absolutely continuous, that is, is possessed of a density. In view of the increasing nature of the d.f. this certainly seems plausible (though the reader familiar with Karl Weierstrass's construction of a continuous function which is nowhere differentiable—see Section X.8—may be more cautious in providing such an endorsement).

EXAMPLE 4) The Cantor distribution. We consider the Cantor set anew. With notation as in Example XI.3.3, the Cantor set \mathbb{C} is the complement of an open set \mathbb{U} of Lebesgue measure one which is contained in the unit interval. It will be convenient to append the open intervals $(-\infty, 0)$ and $(1, \infty)$ to the set \mathbb{U} to form the set $\mathbb{U}_0 = (-\infty, 0) \cup \mathbb{U} \cup (1, \infty)$. If, for each i , we set $\mathbb{J}_{i,0} = (-\infty, 0)$ and $\mathbb{J}_{i,2^i} = (1, \infty)$, we may also write $\mathbb{U}_0 = \bigcup_i \bigcup_j \mathbb{J}_{i,j}$. As \mathbb{U} is open and dense in the unit interval $[0, 1]$ it follows that \mathbb{U}_0 is open and dense in the real line \mathbb{R} .

For each $i \geq 1$ and $0 \leq j \leq 2^i$ we set $c_{i,j} = j/2^i$. We now define a function F_0 on the set \mathbb{U}_0 by setting $F_0(x) = c_{i,j}$ if x is in $\mathbb{J}_{i,j}$. As a point x in \mathbb{U} may lie in many different intervals $\mathbb{J}_{i,j}$ we first need to ensure that our definition is consistent. Now the construction makes clear that two intervals $\mathbb{J}_{i,j}$ and $\mathbb{J}_{i',j'}$ must either coincide or be disjoint. It will suffice hence to show that $c_{i,j} = c_{i',j'}$ whenever the corresponding intervals $\mathbb{J}_{i,j}$ and $\mathbb{J}_{i',j'}$ coincide. This is readily verifiable by an inductive argument if we proceed step by step. For two successive stages we observe that $\mathbb{J}_{i,j} = \mathbb{J}_{i+1,2j}$ and it is readily verifiable that $c_{i,j} = c_{i+1,2j}$. So certainly our assignment of values to the function F_0 is consistent over any two steps. By induction it is consistent over all steps.

A set of points on which a function takes a given constant value is called a *level set* of the function. As $F_0 = c_{i,j}$ over the interval $\mathbb{J}_{i,j}$, the intervals $\mathbb{J}_{i,j}$ (shown through five stages in Table XI.1) constitute the level sets of F_0 . The graph of F_0 plateaus over the level sets $\mathbb{J}_{i,j}$ and the resulting level curves are sketched in Figure 1 through the first four stages of the inductive specification. The reader is invited to imagine an endless procession of a larger and larger number of tinier and tinier level sets ultimately filling up all the interstices between the intervals shown, with the graph of F_0 increasing smoothly from a value of $F_0(x) = 0$ for $x < 0$ to a value of $F_0(x) = 1$ for $x > 1$ in a succession of plateaus with edges at the points of the Cantor set \mathbb{C} of measure zero. The graph of F_0 suggests that we have a distribution function in the making;

to make it a bona fide d.f. we will have to extend F_0 to a function F on the entire real line by specifying values at the points of the Cantor set $C = \mathbb{U}_0^c$. In order to preserve right continuity we are forced to define F , if at all, by $F(x) = \inf\{F_0(t) : x < t, t \in \mathbb{U}_0\}$. As F_0 is increasing on \mathbb{U}_0 and \mathbb{U}_0 is dense in \mathbb{R} it follows immediately that F agrees with F_0 on \mathbb{U}_0 . Thus, F is an extension of F_0 to the real line and it is now apparent that F increases smoothly from 0 to 1 with no jumps, that is to say, it is a continuous d.f. The formal verification of what seems to be suggested graphically may be skipped on a first reading.

The d.f. that we've constructed is called the *Cantor distribution*. That it is continuous everywhere may not raise eyebrows, perhaps, but the reader may find it remarkable that its points of increase are confined to the Cantor set C whose Lebesgue measure is zero. As the intervals $J_{i,j}$ are level sets of F , it follows that F' exists at least over these intervals and is identically zero there. And as the family of intervals $\{J_{i,j}\}$ is dense on the real line it follows that F' is zero almost everywhere and hence cannot represent a density. Thus, *we are confronted with a continuous d.f. F for which a density is abeyant and F hence does not have a representation of the form (4.1)*. ▶

Proof that F is a continuous d.f.: As the intervals $J_{i,0} = (-\infty, 0)$ all coincide, if $x < 0$ we have $F_0(x) = 0/2^i = 0$; likewise, the intervals $J_{i,2^i}$ coincide so that when $x > 1$ we have $F_0(x) = 2^i/2^i = 1$. Now the value of F_0 is constant on a given level set $J_{i,j}$ and is strictly larger on any other level set $J_{i,j'}$ to the right of $J_{i,j}$. Suppose now that x and x' are points of \mathbb{U}_0 and $x < x'$. Then there exist intervals $J_{i,j}$ and $J_{i',j'}$ such that $x \in J_{i,j}$ and $x' \in J_{i',j'}$. We may suppose $i = i'$ by, if necessary, increasing the smaller index of the two and selecting the interval at that level containing the given point. Thus, for some value of the index i , we have $x \in J_{i,j}$ and $x' \in J_{i,j'}$ with $j \leq j'$. But then this means that $F_0(x) \leq F_0(x')$ and F_0 is increasing on \mathbb{U}_0 .

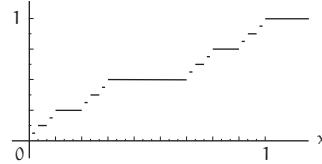


Figure 1: The ternary Cantor distribution.

To verify that the graph of F_0 has no jumps in \mathbb{U}_0 we begin with the observation that, for each i , the interval $J_{i,j}$ is at a distance of 3^{-i} or greater from the other intervals $J_{i,j'}$ at that level. If the intervals $J_{i,0}, \dots, J_{i,2^i}$ are removed then we are left with 2^i disjoint, interstitial intervals, each of length 3^{-i} and over each of which F_0 increases by 2^{-i} . Thus, if x and x' are two points in \mathbb{U}_0 with $|x - x'| < 3^{-i}$ then they are either both in some common interval $J_{i,j}$, or both in an interstice of length 3^{-i} between a pair of the $J_{i,j}$, or one in some $J_{i,j}$ with the other in an adjoining interstitial closed interval of length 3^{-i} . In all cases we have $|F_0(x) - F_0(x')| \leq 2^{-i}$, the upper bound as small as desired for all sufficiently large values of the index i . It follows that F_0 is in fact uniformly continuous, and *a fortiori* right continuous, over \mathbb{U}_0 . We're now ready to consider F .

Suppose $x < x'$. By definition of infimum, for every $\epsilon > 0$, there exists $y \in \mathbb{U}_0$ with $y > x'$ and $F_0(y) - \epsilon < F(x') \leq F_0(y)$. But, as $y > x$, again by definition of infimum, $F(x) \leq F_0(y)$. It follows that $F(x) < F(x') + \epsilon$ and as $\epsilon > 0$ may be chosen arbitrarily small we must have perforce that $F(x) \leq F(x')$ and F is increasing.

Finally, let x and x' be real with $|x - x'| < 3^{-i}$. We may suppose without loss of generality that $x' > x$. Then, by definition of infimum again, there exists $y > x$ with $y \in \mathbb{U}_0$ for which $F_0(y) - \epsilon < F(x) \leq F_0(y)$. Writing $\delta = y - x > 0$, we may now select y' in \mathbb{U}_0 with $x' < y' \leq x' + \delta$ for which $F_0(y') - \epsilon < F(x') \leq F_0(y')$. We observe that $y' - y \leq x' + \delta - y = x' - x < 3^{-i}$. As $F(x') - F(x) = (F(x') - F_0(y')) + (F_0(y') - F_0(y)) + (F_0(y) - F(x))$, two applications of the triangle inequality show that

$$|F(x') - F(x)| \leq |F(x') - F_0(y')| + |F_0(y') - F_0(y)| + |F_0(y) - F(x)| < 2\epsilon + 2^{-i}$$

with the bound on the right less than, say, 3ϵ for all sufficiently large i . It follows that $|F(x') - F(x)| < 3\epsilon$ whenever $|x' - x| < 3^{-i}$ and i is sufficiently large. As $\epsilon > 0$ may be chosen arbitrarily small, F is continuous (and *a fortiori* right continuous). ▶

The Cantor d.f. shows that it is possible to have a continuous d.f. whose derivative (exists and) is zero almost everywhere. This is a perplexing state of affairs, but having unearthed it we will have to deal with it.

DEFINITION 2 A d.f. F is *singular* if F' (exists and) is equal to zero a.e.; F is *singular continuous* if it is continuous and singular.

The reader may find it plausible that any discrete d.f. is singular as the points of increase are countable. (Intuition is served well enough if the graph of the discrete d.f. is comprised of a series of steps, as in the case of simple random variables. The reader should pause to consider Example 3.6, however, to see that even in the discrete case the analysis of the derivative is not, in general, completely trivial.) The Cantor d.f. highlights the remarkable fact that even continuity is not a bar to singularity. The curious reader will see in the *Problems* that no further variations of elementary d.f.s need be considered.

5 Modes of convergence

A key feature that makes the concept of measurability so useful is that limiting operations on measurable functions preserve measurability. To obviate the necessity of frequently making caveats about the values $\pm\infty$ when dealing with possibly divergent sequences of functions it will be useful to expand the definition of measurability to functions with divergent values at individual points.

Let $\bar{\mathbb{R}}$ be the extended real line obtained by appending the elements $-\infty$ and $+\infty$ to \mathbb{R} . The corresponding extended Borel σ -algebra $\bar{\mathcal{B}}$ consists of the Borel sets in \mathcal{B} , possibly enlarged by addition of one or both of the points $\pm\infty$. The measurability theorem of Section 1 carries through *in toto* if X is an extended real-valued, measurable function and the sets \mathbb{I} are identified with extended Borel sets.

THEOREM 1 Suppose $\{X_n, n \geq 1\}$ is a sequence of extended real-valued functions measurable with respect to some σ -algebra \mathcal{F} . Then $\sup X_n$, $\inf X_n$, $\limsup X_n$, and $\liminf X_n$ are all measurable with respect to \mathcal{F} , if possibly extended real-valued.

PROOF: As $\inf X_n = -\sup(-X_n)$ and $\liminf X_n = -\limsup(-X_n)$, it will suffice to show that $\sup X_n$ and $\limsup X_n$ are measurable. Now $\sup X_n$ always exists, if possibly extended real-valued. As, for each x , $X_n^{-1}(-\infty, x] \in \mathcal{F}$ via the measurability theorem of Section 1, it follows that $(\sup X_n)^{-1}(-\infty, x] = \bigcap_n X_n^{-1}(-\infty, x]$ is also an element of \mathcal{F} . Consequently,

$$(\sup X_n)^{-1}(a, b] = ((\sup X_n)^{-1})(-\infty, b]) \setminus ((\sup X_n)^{-1})(-\infty, a]) \in \mathcal{F}$$

for each half-closed interval $(a, b]$, and it follows that $\sup X_n$, hence also $\inf X_n$, are measurable. To finish off the proof, it suffices to observe that $\limsup X_n = \inf_{n \geq 1} (\sup_{m \geq n} X_m)$, and, as by the just concluded argument $Y_n = \sup_{m \geq n} X_m$ is measurable, so is $\inf_n Y_n$ by another application of the argument. ►

As an immediate consequence, if $\{X_n\}$ is an ordinary sequence of measurable real-valued functions converging pointwise to a real-valued function X then $X = \lim X_n = \limsup X_n = \liminf X_n$ is measurable. It follows that *the family of measurable functions [on any given measurable space (Ω, \mathcal{F})] is closed under pointwise limits*. If we equip the measurable space (Ω, \mathcal{F}) with a probability measure \mathbf{P} then we conclude that *any convergent sequence $\{X_n\}$ of random variables has a limit $X = \lim X_n$ that is also a random variable*. If X is permitted to take values $\pm\infty$, that is to say, it is an extended real-valued measurable function, we refer to it as a *defective random variable* with associated *defective distribution* which places mass at $\pm\infty$. To avoid terminological boondoggles we will reserve the term *random variable* in ordinary discourse to mean ordinary measurable real-valued functions on a probability space and use the modifier *defective* when necessary to accommodate the values $\pm\infty$. If the sequence $\{X_n\}$ is not convergent everywhere then by focusing on the set $\Omega_0 \in \mathcal{F}$ of sample points on which $\limsup X_n = \liminf X_n$, we may conclude that $\lim_n X_n$ is a (possibly defective) random variable on Ω_0 .

Sequences of random variables that fail to be convergent but only on a negligible set form an important part of the story. In analogy with the corresponding notion of sets of Lebesgue measure zero on the real line, when we are dealing with an abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$ we say that a property holds *almost everywhere (a.e.)* if it holds for all sample points ω excepting only on a *null set* \mathfrak{N} of aberrant points whose measure is zero, that is to say, $\mathbf{P}(\mathfrak{N}) = 0$. We may now talk about the (pointwise) convergence of a sequence of random variables $X_n(\omega)$ excepting, perhaps, only on an irritating null set.

DEFINITION 1 A sequence of random variables $\{X_n, n \geq 1\}$ converges almost everywhere to a random variable X , in notation, $X_n \rightarrow^{\text{a.e.}} X$, if $\mathbf{P}\{\omega : X_n(\omega) \not\rightarrow X(\omega)\} = 0$.

Convergence almost everywhere (a.e.) is also known as convergence *almost surely* (a.s.) or convergence *almost certainly* (a.c.) or convergence *with probability one* (w.p.1.). We have seen an example of convergence almost everywhere in Borel's

law of normal numbers in Section V.7: if S_n represents the accumulated number of successes in n tosses of a fair coin then $\frac{1}{n}S_n \rightarrow^{\text{a.e.}} \frac{1}{2}$.

In dealing with a sequence of random variables the reader should learn to reason with a single sample point, say, ω , and the corresponding sequence of real values $\{X_n(\omega), n \geq 1\}$, and then translate the intuition garnered about the sequence into probabilistic statements. This way of thinking is illustrated in the following characterisation of convergence a.e. We recall that, mirroring notation for real sequences, if $\{A_n\}$ is a sequence of sets then $B_n = \bigcap_{m \geq n} A_m$ is an increasing sequence of sets and we write $\limsup_n A_n = \lim_n B_n = \bigcup_n B_n$; likewise, $C_n = \bigcup_{m \geq n} A_m$ is a decreasing sequence and we write $\liminf_n A_n = \lim_n C_n = \bigcap_n C_n$.

THEOREM 2 *A random sequence $\{X_n, n \geq 1\}$ converges a.e. to a random variable X if, and only if, for every $\epsilon > 0$, we have $P(\limsup_n \{|X_n - X| < \epsilon\}) = 1$ or, equivalently, $P(\liminf_n \{|X_n - X| \geq \epsilon\}) = 0$.*

PROOF: Let Ω_0 denote the set of sample points ω on which $X_n(\omega) \rightarrow X(\omega)$ and, for any $\epsilon > 0$, write $A_n(\epsilon)$ for the event $\{|X_n - X| < \epsilon\}$. Suppose first that $X_n \rightarrow^{\text{a.e.}} X$. If $\omega \in \Omega_0$ then ω is in $A_n(\epsilon)$ eventually (for all sufficiently large n), and hence also in $\limsup_n A_n(\epsilon)$. Thus Ω_0 is a subset of $\limsup_n A_n(\epsilon)$ and it follows that $1 = P(\Omega_0) \leq P(\limsup_n A_n(\epsilon)) \leq 1$.

To argue the converse, suppose $P(\limsup_n A_n(\epsilon)) = 1$ for every $\epsilon > 0$. Select any sequence of values $\epsilon = \epsilon_k$ decreasing monotonically to zero, say, $\epsilon_k = 1/k$. The sets $\limsup_n A_n(1/k)$ each have probability one and, as $k \rightarrow \infty$, decrease monotonically to a limit set A . It follows by continuity of probability measure that $P(A) = 1$. Suppose $\omega \in A$. Then $|X_n(\omega) - X(\omega)| < \epsilon$ eventually for all $\epsilon = 1/k$, hence for all $\epsilon > 0$, whence $\omega \in \Omega_0$. It follows that $P(\Omega_0) = 1$ and hence $X_n \rightarrow X$ a.e. ▶

When a limiting random variable is not clearly in view it will suffice to show that the sequence is Cauchy.

THEOREM 3 *A random sequence $\{X_n, n \geq 1\}$ converges a.e. if, and only if, for every $\epsilon > 0$, as $n \rightarrow \infty$, we have $P(\bigcup_{j \geq 1} \{|X_{n+j} - X_n| \geq \epsilon\}) \rightarrow 0$.*

PROOF: It will be enough to prove sufficiency; necessity follows as before. Write $B_n(\epsilon) = \bigcup_{j,k \geq n} \{|X_j - X_k| \geq \epsilon\}$. The triangle inequality $|X_j - X_k| \leq |X_j - X_n| + |X_k - X_n|$ shows that the occurrence of $\{|X_j - X_k| \geq \epsilon\}$ implies the occurrence of at least one of the events $\{|X_j - X_n| \geq \epsilon/2\}$ and $\{|X_k - X_n| \geq \epsilon/2\}$. Under the conditions of the theorem we hence have $P(B_n(\epsilon)) \rightarrow 0$ as $n \rightarrow \infty$. As $B_n(\epsilon)$ decreases to the limit set $B(\epsilon) = \bigcap_n B_n(\epsilon)$, by continuity of probability measure it follows that $P(B(\epsilon)) = 0$ for every $\epsilon > 0$. By allowing ϵ to decrease to zero through rational points, say, the sequence $\{1/m, m \geq 1\}$, the

sets $B(1/m)$ increase to a limit set B . Then $\mathbf{P}(B) = 0$ again by continuity of probability measure. If, for some ω , the sequence $\{X_n(\omega), n \geq 1\}$ is not Cauchy then there exists $m > 0$ such that $|X_j(\omega) - X_k(\omega)| \geq 1/m$ i.o. This means that ω is an element of $B_n(1/m)$ for each n and is hence also in $B(1/m)$. It follows that the set of sample points ω for which $\{X_n(\omega), n \geq 1\}$ is not Cauchy is contained in the probability zero set B . ►

Convergence a.e. frequently represents the gold standard for convergence but is a demanding concept. The following weaker form of convergence is easier to show in many applications and is hence of basic importance.

DEFINITION 2 A sequence of random variables $\{X_n, n \geq 1\}$ converges in probability to a random variable X , in notation, $X_n \rightarrow^P X$, if $\lim_n \mathbf{P}\{|X_n - X| \geq \epsilon\} = 0$ for every $\epsilon > 0$.

Convergence in probability is also called convergence *in measure*. The weak law of large numbers in Section V.7 provides an example: $\frac{1}{n} S_n \rightarrow^P \frac{1}{2}$. In view of Theorem 2, convergence in probability is weaker than convergence a.e.

THEOREM 4 *Convergence a.e. implies convergence in probability.*

The converse implication will hold provided the tails of the distribution of $X_n - X$ die sufficiently quickly. The simplest case is suggested by the Borel–Cantelli lemma of Section IV.4. If $\sum_n \mathbf{P}\{|X_n - X| \geq \epsilon\} < \infty$ then $\mathbf{P}\{|X_n - X| \geq \epsilon \text{ i.o.}\} = 0$, or, equivalently, only finitely many of the events $\{|X_n - X| > \epsilon\}$ occur. But this is the same as saying $\lim_n \mathbf{P}(\bigcap_{m \geq n} \{|X_m - X| < \epsilon\}) = 1$. A standard way of showing that a sequence converges a.e. is hence to begin by showing that it converges in probability and then leveraging the Borel–Cantelli lemma to show that the convergence is in fact a.e.

6 Baire functions, coordinate transformations

Suppose X is a random variable on an abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We think of X as playing the rôle of a coordinate variable. We would like to consider functions $Y = g(X)$ of the coordinate variable as forming new random variables from old. What sort of functions should be admissible? Well, continuous functions appear to be a given. We will need to verify, however, that the composition $Y(\omega) = g(X(\omega))$ is indeed a random variable, that is to say, a measurable map on the space $(\Omega, \mathcal{F}, \mathbf{P})$.

While measurability is a somewhat opaque concept in abstract, geometric intuition is strong when dealing with functions on a Euclidean space. This is the only place in the narrative where we will need the technical results from Section I.9 and the reader should glance at this material before proceeding.

THEOREM 1 Suppose X is a measurable map on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and g a continuous function of a real variable. Then the composition $Y(\omega) = (g \circ X)(\omega) = g(X(\omega))$ is also a measurable map (that is to say, a random variable) on $(\Omega, \mathcal{F}, \mathbf{P})$.

PROOF: It will suffice to show that g is Borel-measurable. For then, if $(a, b]$ is any half-closed interval, its preimage $g^{-1}(a, b]$ is a Borel set. But X is a bona fide random variable hence \mathcal{F} -measurable. The preimage under X of the Borel subset $g^{-1}(a, b]$ of the line is hence an element of \mathcal{F} .

By the definition of continuity, if x_0 is any real number and $\epsilon > 0$ then there exists $\delta > 0$ determined by x_0 and ϵ so that $|g(x) - g(x_0)| < \epsilon$ whenever $|x - x_0| < \delta$. Suppose now that (a, b) is any open interval. We claim then that the preimage under g of (a, b) is an open set, that is to say, for any point x_0 in $g^{-1}(a, b)$ there exists $\delta > 0$ such that the open interval $(x_0 - \delta, x_0 + \delta)$ is wholly contained in $g^{-1}(a, b)$. Indeed, suppose x_0 is any point in $g^{-1}(a, b)$. Then $a < y_0 = g(x_0) < b$ and we select for the strictly positive ϵ a sufficiently small quantity so that the open interval $(y_0 - \epsilon, y_0 + \epsilon)$ is contained within (a, b) . But then this implies that the open interval $x_0 - \delta < x < x_0 + \delta$ is contained within the set $g^{-1}(a, b)$. It follows that the preimage under g of (a, b) is an open set and hence also a Borel set as any open set on the line may be represented as the union of a countable number of disjoint open intervals by Theorem I.9.1.

As the half-closed interval $(a, b]$ is the intersection of a countable number of open intervals $(a, b + 1/n)$, it follows that $g^{-1}(a, b] = g^{-1}(\bigcap_n (a, b + 1/n)) = \bigcap_n g^{-1}(a, b + 1/n)$ is the intersection of a countable number of Borel sets on the line, hence also a Borel set. Thus g is Borel-measurable and the composite function $Y = g(X)$ is indeed measurable with respect to \mathcal{F} . ▶

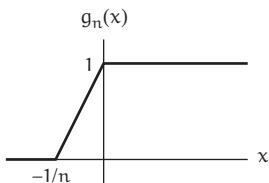


Figure 2: Convergence to the Heaviside function.

The continuous functions are not nearly rich enough for our purposes as one can imagine a whole raft of otherwise reasonable functions that are not continuous. The next step may be to include functions with jump discontinuities. It may occur to the reader that any such function g may be represented as the pointwise limit of a sequence of continuous functions $\{g_n\}$. As an illustration, the sequence of continuous functions $\{g_n(x), n \geq 1\}$ in Figure 2 converges pointwise to the Heaviside function $H_0(x)$.

As the only effective procedure of constructing essentially new functions is via limiting operations, we are hence led to consider functions in a class including the continuous functions and closed under pointwise limits. There is at least one such class, namely, the class of all functions. The intersection of all such families of functions is the smallest class that includes the continuous functions and is closed under pointwise limits. Prudence leads us to consider this class as the least complicated collection of functions containing enough variety for our purposes.

DEFINITION The *Baire class* \mathfrak{B} of functions is the smallest collection of functions that includes the continuous functions and is closed under pointwise limits. In other words, if g is continuous then $g \in \mathfrak{B}$ and if $\{g_n\}$ is a sequence of functions in \mathfrak{B} with $g_n \rightarrow g$ pointwise then $g \in \mathfrak{B}$. We say that a member of this class is a *Baire function*. We shall also use this terminology for functions whose domain is restricted to a subset of the reals, say, the positive half-line, and also for functions of more than one variable.

By iterations of the procedure shown in Figure 2 it is clear that functions whose only discontinuities are jumps at a finite, or even a countable, number of locations are Baire functions (and *a fortiori* all distribution functions are Baire functions). Limits of such sequences of functions lead to more complicated Baire functions, limits of such functions leading to yet more complicated functions, and proceeding in this fashion a satisfyingly rich family of functions is built up which serves all the needs of elementary probability.

Suppose X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and g is any (univariate) Baire function. We are then led to consider the measurability of the composition $Y = g \circ X$.

THEOREM 2 *Every Baire function is Borel-measurable.*

PROOF: It will suffice to show that g is Borel-measurable. By Theorem 1, the class of Borel-measurable functions contains all continuous functions and, by Theorem 5.1, is closed under pointwise limits. But the Baire class is the *smallest* class of functions that contains the continuous functions and is closed under pointwise limits. It follows that all Baire functions are Borel-measurable. ►

The converse is also true in Euclidean spaces, that is to say, the class of Borel-measurable functions coincides with the Baire class on \mathbb{R}^n , but we will not stop to pull this particular chestnut out of the fire.

Thus, ordinary operations on a random variable X lead to new random variables, $H_0(X)$, X^n , $\sin(X)$, e^X , and $\log(X)$, for instance, are all random variables. In general, $Y = g(X)$ is a random variable for any Baire function g , the measure induced by Y on the Borel sets of the line implicitly determined by that of X . Indeed, if X and Y have d.f.s F and G , respectively, it is clear that we may write $G(\mathbb{I}) = F(g^{-1}\mathbb{I})$ for every Borel set \mathbb{I} .

7 Two and more dimensions

Our considerations for individual random variables carry over naturally to two and more variables. I will focus on the setting in two dimensions as it captures the essential features. Suppose X_1 and X_2 are random variables defined on an abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then the pair (X_1, X_2) determines a map

from Ω into the Euclidean plane \mathbb{R}^2 . If \mathbb{I}_1 and \mathbb{I}_2 are half-closed intervals on the line then

$$(X_1, X_2)^{-1}(\mathbb{I}_1 \times \mathbb{I}_2) = \{\omega : X_1(\omega) \in \mathbb{I}_1, X_2(\omega) \in \mathbb{I}_2\} = X_1^{-1}\mathbb{I}_1 \cap X_2^{-1}\mathbb{I}_2.$$

It follows that the preimage under (X_1, X_2) of the half-closed rectangle $\mathbb{I}_1 \times \mathbb{I}_2$ is an event, the argument remaining impeccable if we replace the half-closed intervals \mathbb{I}_1 and \mathbb{I}_2 by Borel sets on the line by virtue of the measurability theorem of Section 1. We run into the difficulty, however, that while Cartesian products of Borel sets on the line certainly yield Borel sets on the plane, a generic Borel set in \mathbb{R}^2 need not be representable as a Cartesian product of Borel sets in \mathbb{R}^1 . We hence proceed indirectly.

Let \mathcal{G} be the family of subsets \mathbb{A} of \mathbb{R}^2 for which $(X_1, X_2)^{-1}\mathbb{A} \in \mathcal{F}$. Then \mathcal{G} certainly contains the half-closed rectangles of the form $(a_1, b_1] \times (a_2, b_2]$ and, more generally, contains the Cartesian product $\mathbb{I}_1 \times \mathbb{I}_2$ of any two Borel sets \mathbb{I}_1 and \mathbb{I}_2 of the line. The steps now follow the proof of the measurability theorem almost verbatim—little more is required than to strike out the odd reference to \mathbb{R} and replace it by \mathbb{R}^2 —to conclude that \mathcal{G} is a σ -algebra of subsets of the plane containing the family of half-closed rectangles. As the Borel σ -algebra $\mathcal{B}(\mathbb{R}^2)$ of subsets of the Euclidean plane \mathbb{R}^2 is the smallest σ -algebra containing the half-closed rectangles, it must perforce be contained in \mathcal{G} , and we obtain the following two-dimensional version of the measurability theorem.

THEOREM 1 *Suppose X_1 and X_2 are random variables on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Then $(X_1, X_2)^{-1}\mathbb{A} \in \mathcal{F}$ for every Borel set \mathbb{A} of the Euclidean plane \mathbb{R}^2 .*

It follows that to each Borel set \mathbb{A} in \mathbb{R}^2 we may naturally associate the probability $\mathbf{P}((X_1, X_2)^{-1}\mathbb{A})$. In particular, to each half-closed rectangle $(a_1, b_1] \times (a_2, b_2]$ we may associate the probability $\mathbf{P}(X_1^{-1}(a_1, b_1] \cap X_2^{-1}(a_2, b_2])$. In analogy with terminology in one dimension, we call the pair (X_1, X_2) a random variable in two dimensions (or *random vector*) and we may now define the *two-dimensional d.f.* F of the pair (X_1, X_2) by

$$\begin{aligned} F(x_1, x_2) &= \mathbf{P}((X_1, X_2)^{-1}(-\infty, x_1] \times (-\infty, x_2]) \\ &= \mathbf{P}\{\omega : X_1(\omega) \leq x_1, X_2(\omega) \leq x_2\} = \mathbf{P}\{X_1 \leq x_1, X_2 \leq x_2\}, \end{aligned}$$

the expressions all being just notational variants of one another. As one might anticipate, F is increasing in each argument and satisfies

$$\begin{aligned} \lim_{x_1 \rightarrow -\infty} F(x_1, x_2) &= F(-\infty, x_2) = 0, & \lim_{x_2 \rightarrow -\infty} F(x_1, x_2) &= F(x_1, -\infty) = 0, \\ \lim_{\substack{x_1 \rightarrow \infty \\ x_2 \rightarrow \infty}} F(x_1, x_2) &= F(+\infty, +\infty) = 1. \end{aligned}$$

Moreover, $\lim_{x_2 \rightarrow \infty} F(x_1, x_2) = F_1(x_1)$ and $\lim_{x_1 \rightarrow \infty} F(x_1, x_2) = F_2(x_2)$ are just the (marginal or one-dimensional) d.f.s of X_1 and X_2 , respectively. I will leave the reader to verify these properties.

The probabilities of half-closed rectangles may be expressed in terms of the d.f. though the expressions are now more cumbrous than in the one-dimensional case. Additivity of probability measure yields the expression

$$P\{a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2\} = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2). \quad (7.1)$$

A glance at Figure 3 will make the logic clear; the reader may recognise an inclusion–exclusion argument! Carathéodory's extension theorem works seamlessly in this setting as well and the d.f. F of the pair (X_1, X_2) induces a unique probability measure on the Borel sets of \mathbb{R}^2 which satisfies (7.1) on the half-closed rectangles. (The interested reader should leaf through the proof for the one-dimensional case and convince herself that nothing material is changed in moving to more than one dimension.) In keeping with our notational conventions for one-dimensional d.f.s we simply write F for this probability measure as well, the argument serving to establish whether F is in its rôle of a d.f. in two dimensions or as a probability measure on the Borel sets of the plane. Thus, $F((a_1, b_1] \times (a_2, b_2])$ stands for the expression on the right of (7.1) and, in general, for any Borel set \mathbb{A} in the plane $F(\mathbb{A}) = P\{(X_1, X_2)^{-1}\mathbb{A}\}$ stands for the probability that the pair (X_1, X_2) takes a value in \mathbb{A} . As in the one-dimensional case, all probabilistic questions about the random pair (X_1, X_2) can be resolved by the d.f. $F(x_1, x_2)$, equivalently the probability measure $F(\mathbb{A})$ on \mathbb{R}^2 , without the necessity to refer back to the engendering probability measure P in the original space Ω .

New random variables are obtained by coordinate transformations on the plane.

THEOREM 2 Suppose X_1 and X_2 are random variables on an abstract probability space. Let $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ be any Baire function of two real variables. Then $g(X_1, X_2)$ is a random variable defined on the same space as X_1 and X_2 .

PROOF: Let \mathbb{I} be any half-closed interval of the line. Then $g(X_1, X_2)^{-1}\mathbb{I} = (X_1, X_2)^{-1}(g^{-1}\mathbb{I})$ and, in view of Theorem 1, it suffices to show that $g^{-1}\mathbb{I}$ is a Borel subset of the Euclidean plane \mathbb{R}^2 . The rest of the proof follows the pattern for one dimension and I will leave it to the reader to supply the details. ►

It follows *a fortiori* that ordinary arithmetic operations yield random variables. Thus, $\max\{X_1, X_2\}$, $\min\{X_1, X_2\}$, $X_1 + X_2$, $X_1 - X_2$, $X_1 \cdot X_2$, and X_1/X_2 (the last if X_2 is non-zero) are all random variables. In general, if the pair (X_1, X_2) has d.f. $F(x_1, x_2)$, and $Y = g(X_1, X_2)$ has d.f. $G(y)$, then the induced measure of Y is given by $G(\mathbb{I}) = F((X_1, X_2)^{-1}\mathbb{I})$ for each Borel set \mathbb{I} of the line.

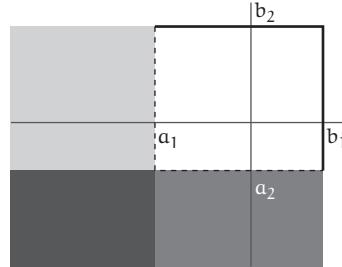


Figure 3: Inclusion–exclusion.

It is clear that the arguments of this section are not specialised to only two dimensions and indeed carry over to any number of dimensions; we accordingly simply note the multi-dimensional version of Theorem 1, the proof following that for two dimensions verbatim.

THEOREM 1' Suppose X_1, \dots, X_n are random variables with respect to some σ -algebra \mathcal{F} . Then $(X_1, \dots, X_n)^{-1}\mathbb{A} \in \mathcal{F}$ for every Borel set \mathbb{A} of n -dimensional Euclidean space \mathbb{R}^n .

In other words, the map $\omega \mapsto (X_1, \dots, X_n)(\omega)$ is an \mathcal{F} -measurable vector-valued function of the sample space Ω . We naturally identify $\mathbf{X} = (X_1, \dots, X_n)$ as a random variable in n dimensions (or, to emphasise the dimensionality, a random *vector*). To each Borel set \mathbb{A} of \mathbb{R}^n we may now associate the probability $\mathbf{P}\{(X_1, \dots, X_n)^{-1}\mathbb{A}\}$ and the vector random variable (X_1, \dots, X_n) induces a unique probability measure on the Borel sets of \mathbb{R}^n . As is the case in one and two dimensions, we may characterise this measure by a consideration of rectangles.

The n -dimensional d.f. of (X_1, \dots, X_n) is the function

$$\begin{aligned} F(x_1, \dots, x_n) &= \mathbf{P}((X_1, \dots, X_n)^{-1}(-\infty, x_1] \times \dots \times (-\infty, x_n]) \\ &= \mathbf{P}\{\omega : X_1(\omega) \leq x_1, \dots, X_n(\omega) \leq x_n\} = \mathbf{P}\{X_1 \leq x_1, \dots, X_n \leq x_n\}. \end{aligned}$$

The d.f. determines via the extension theorem a unique probability measure on the Borel sets of Euclidean space \mathbb{R}^n . The n -dimensional expressions corresponding to (7.1) are now unfortunately much more cumbersome to write down and in two and more dimensions the d.f. loses much of the elegance it has in one dimension; in particular, unlike the case in one dimension, monotonicity in each argument does not suffice to make a properly normalised, positive function in two or more dimensions a d.f.—see Problem XI.1. Consequently, in higher dimensions it is usually more convenient to work directly with the induced measure $F(\mathbb{A})$ on the Borel sets \mathbb{A} of \mathbb{R}^n rather than with the corresponding d.f. $F(x_1, \dots, x_n)$.

The reader has seen simple examples of distributions in two and more dimensions scattered through Chapters VII–X.

As in one and two dimensions, new random variables may be generated by suitable coordinate transformations on \mathbb{R}^n . We simply remark the following extension of Theorem 2, the proof being almost identical.

THEOREM 2' Suppose X_1, \dots, X_n are random variables on some abstract probability space and $g: \mathbb{R}^n \rightarrow \mathbb{R}$ is a Baire function of n variables. Then $Y = g(X_1, \dots, X_n)$ is a random variable defined on the same space as X_1, \dots, X_n .

In particular, if $F(x_1, \dots, x_n)$ is the n -dimensional d.f. of (X_1, \dots, X_n) and $G(y)$ is the d.f. of Y then the induced measure for Y is given by $G(\mathbb{I}) = F(g^{-1}\mathbb{I})$ for every Borel set \mathbb{I} of the line.

8 Independence, product measures

We now return to the fundamental concept of independence so peculiar to the theory of probability and that is responsible for so much of its intuitive character. A variety of vantage points and applications of independence that the reader has thus far been exposed to has served to build intuition; we are now equipped to proceed to a rather general formulation of the concept. The case of two random variables provides a template for the general case and it is simplest to begin with this setting. We assume that all random variables are defined on a common underlying probability space.

DEFINITION 1 Random variables X_1 and X_2 are *independent* if, and only if,

$$\mathbf{P}\{X_1 \in \mathbb{A}_1, X_2 \in \mathbb{A}_2\} = \mathbf{P}\{X_1 \in \mathbb{A}_1\} \mathbf{P}\{X_2 \in \mathbb{A}_2\} \quad (8.1)$$

for all Borel sets \mathbb{A}_1 and \mathbb{A}_2 of the real line.

The reader who has read Section III.5 will realise that no new ground has been covered: our definition is equivalent to the statement that the generated σ -algebras $\sigma(X_1)$ and $\sigma(X_2)$ are independent; when cast explicitly in terms of the constituent random variables themselves, however, the definition has a more intuitive appeal.

The reader may feel that the definition is natural in that it extends the rule of products in Definition III.1.1 to the intersection of any pair of events (Borel sets) but she may well feel a niggling worry that the definition places very strong constraints on the variables. After all there are a very large number of events (Borel sets) on the line and the definition requires a rule of products for *any* pair of them. Can all these constraints be simultaneously satisfied by a probability measure \mathbf{P} ? An alternative viewpoint helps reassure on that score.

Suppose the random pair (X_1, X_2) is possessed of the two-dimensional d.f. $F(x_1, x_2)$, the individual variables X_1 and X_2 having the corresponding marginal d.f.s $F_1(x_1)$ and $F_2(x_2)$, respectively. In terms of the induced measures F on the plane and F_1, F_2 on the line, we may restate the defining relation (8.1) in the form

$$F(\mathbb{A}_1 \times \mathbb{A}_2) = F_1(\mathbb{A}_1)F_2(\mathbb{A}_2) \quad (\mathbb{A}_1, \mathbb{A}_2 \in \mathcal{B}). \quad (8.1')$$

By restricting \mathbb{A}_1 and \mathbb{A}_2 to half-closed intervals it now follows *a fortiori* that

$$F(x_1, x_2) = F((-\infty, x_1] \times (-\infty, x_2]) = F_1(-\infty, x_1]F_2(-\infty, x_2] = F_1(x_1)F_2(x_2) \quad (8.1'')$$

where we have cheerfully toggled between the rôles of d.f. and measure. It follows that independence implies a rule of products for rectangles in the plane, hence for the two-dimensional d.f. It may come as a mild surprise that the apparently much more restrictive condition (8.1') is implied by the apparently weaker condition (8.1'').

THEOREM 1 *The random variables X_1 and X_2 are independent if, and only if, the two-dimensional d.f. F may be separated into a product of marginal d.f.s $F(x_1, x_2) = F_1(x_1)F_2(x_2)$.*

PROOF: The demonstration follows a usual pattern in the theory of measure: we first establish the result for intervals, then for finite unions of intervals, and finally for general Borel sets. (The reader who finds first-principle proofs tedious should skip to the end of the proof where she will find a compact alternative.) Since one direction is obvious, we suppose that $F(x_1, x_2) = F_1(x_1)F_2(x_2)$. We first establish (8.1') when \mathbb{A}_1 and \mathbb{A}_2 are elements of the ring $R(\mathcal{I})$ of finite unions of half-closed intervals of the form $(a, b]$. The base case when \mathbb{A}_1 and \mathbb{A}_2 are themselves half-closed intervals follows directly from (7.1):

$$\begin{aligned} F((a_1, b_1] \times (a_2, b_2)) &= F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2) \\ &= F_1(b_1)F_2(b_2) - F_1(b_1)F_2(a_2) - F_1(a_1)F_2(b_2) + F_1(a_1)F_2(a_2) \\ &= (F_1(b_1) - F_1(a_1))(F_2(b_2) - F_2(a_2)) = F_1(a_1, b_1)F_2(a_2, b_2). \end{aligned}$$

We now proceed by a two-stage induction. With $\mathbb{A}_2 = (a_2, b_2]$ any half-closed interval, suppose (8.1') holds when \mathbb{A}_1 is any union of n disjoint half-closed intervals. Suppose $(a_1, b_1]$ is disjoint from \mathbb{A}_1 and consider $\mathbb{A}'_1 = (a_1, b_1] \cup \mathbb{A}_1$. Then

$$\begin{aligned} F(\mathbb{A}'_1 \times (a_2, b_2]) &= F(((a_1, b_1] \times (a_2, b_2]) \cup (\mathbb{A}_1 \times (a_2, b_2])) \\ &= F((a_1, b_1] \times (a_2, b_2)) + F(\mathbb{A}_1 \times (a_2, b_2]) \quad (\text{by additivity}) \\ &= F_1(a_1, b_1]F_2(a_2, b_2) + F_1(\mathbb{A}_1)F_2(a_2, b_2] \quad (\text{by induction hypothesis}) \\ &= (F_1(a_1, b_1] + F_1(\mathbb{A}_1))F_2(a_2, b_2] = F_1(\mathbb{A}'_1)F_2(a_2, b_2] \quad (\text{by additivity}). \end{aligned}$$

As any element of $R(\mathcal{I})$ may be expressed as the union of a finite number of *disjoint* half-closed intervals, by induction, (8.1') holds whenever \mathbb{A}_1 is any element of $R(\mathcal{I})$ and \mathbb{A}_2 is any half-closed interval. We now fix any $\mathbb{A}_1 \in R(\mathcal{I})$ and repeat the induction by allowing \mathbb{A}_2 to be a finite union of half-closed intervals. Accordingly, suppose (8.1') holds when \mathbb{A}_2 is a union of n half-closed intervals disjoint from each other. Suppose $(a_2, b_2]$ is disjoint from \mathbb{A}_2 and let $\mathbb{A}'_2 = (a_2, b_2] \cup \mathbb{A}_2$. Then

$$\begin{aligned} F(\mathbb{A}_1 \times \mathbb{A}'_2) &= F((\mathbb{A}_1 \times (a_2, b_2]) \cup (\mathbb{A}_1 \times \mathbb{A}_2)) = F(\mathbb{A}_1 \times (a_2, b_2]) + F(\mathbb{A}_1 \times \mathbb{A}_2) \\ &= F_1(\mathbb{A}_1)F_2(a_2, b_2] + F_1(\mathbb{A}_1)F_2(\mathbb{A}_2) = F_1(\mathbb{A}_1)(F_2(a_2, b_2] + F_2(\mathbb{A}_2)) = F_1(\mathbb{A}_1)F_2(\mathbb{A}'_2). \end{aligned}$$

The reader should be able to fill in the reasoning paralleling those of the first induction step. This completes the induction. It follows that (8.1') holds whenever \mathbb{A}_1 and \mathbb{A}_2 are finite unions of half-closed intervals. We complete the proof by a limiting argument.

Suppose \mathbb{A}_1 and \mathbb{A}_2 are Borel sets, $\epsilon > 0$ fixed but arbitrary. There then exist $\mathbb{B}_1, \mathbb{B}_2 \in R(\mathcal{I})$ such that $F_1(\mathbb{A}_1 \Delta \mathbb{B}_1) < \epsilon$ and $F_2(\mathbb{A}_2 \Delta \mathbb{B}_2) < \epsilon$. (This is by the very construction of measure from outer measure; see Definitions XI.5.1,2,3. This is also contained in Problem XI.12.) For $j = 1, 2$, we may write $\mathbb{A}_j \cup \mathbb{B}_j = \mathbb{A}_j \cup (\mathbb{B}_j \setminus \mathbb{A}_j) = \mathbb{B}_j \cup (\mathbb{A}_j \setminus \mathbb{B}_j)$ as the union of disjoint sets in two different ways. Thus, on the one hand, we have

$$F((\mathbb{A}_1 \cup \mathbb{B}_1) \times (\mathbb{A}_2 \cup \mathbb{B}_2)) = F(\mathbb{A}_1 \times \mathbb{A}_2) + F(\mathbb{A}_1 \times \mathbb{B}_2 \setminus \mathbb{A}_2) + F(\mathbb{B}_1 \setminus \mathbb{A}_1 \times \mathbb{A}_2) + F(\mathbb{B}_1 \setminus \mathbb{A}_1 \times \mathbb{B}_2 \setminus \mathbb{A}_2)$$

by additivity. The final three terms on the right may be bounded by monotonicity of probability measure by $F(\mathbb{A}_1 \times \mathbb{B}_2 \setminus \mathbb{A}_2) \leq F_2(\mathbb{B}_2 \setminus \mathbb{A}_2) < \epsilon$, $F(\mathbb{B}_1 \setminus \mathbb{A}_1 \times \mathbb{A}_2) \leq F_1(\mathbb{B}_1 \setminus \mathbb{A}_1) <$

ϵ , and $F(\mathbb{B}_1 \setminus \mathbb{A}_1 \times \mathbb{B}_2 \setminus \mathbb{A}_2) \leq F_1(\mathbb{B}_1 \setminus \mathbb{A}_1) < \epsilon$, and it follows that $0 \leq F((\mathbb{A}_1 \cup \mathbb{B}_1) \times (\mathbb{A}_2 \cup \mathbb{B}_2)) - F(\mathbb{A}_1 \times \mathbb{A}_2) < 3\epsilon$. On the other hand, an entirely similar argument using the alternative decomposition of $\mathbb{A}_j \cup \mathbb{B}_j$ shows that $0 \leq F((\mathbb{A}_1 \cup \mathbb{B}_1) \times (\mathbb{A}_2 \cup \mathbb{B}_2)) - F(\mathbb{B}_1 \times \mathbb{B}_2) < 3\epsilon$ also. As \mathbb{B}_1 and \mathbb{B}_2 are finite unions of half-closed intervals we have moreover that $F(\mathbb{B}_1 \times \mathbb{B}_2) = F_1(\mathbb{B}_1)F_2(\mathbb{B}_2)$ and by an application of the triangle inequality it follows that $|F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{A}_1)F_2(\mathbb{B}_2)| < 6\epsilon$. (We may replace “6” by “3” in the upper bound but we need not be overly fussy here.) By grouping terms, we have

$$\begin{aligned} F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{A}_1)F_2(\mathbb{A}_2) &= [F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{B}_1)F_2(\mathbb{B}_2)] \\ &\quad + [F_1(\mathbb{B}_1)F_2(\mathbb{B}_2) - F_1(\mathbb{B}_1)F_2(\mathbb{A}_2)] + [F_1(\mathbb{B}_1)F_2(\mathbb{A}_2) - F_1(\mathbb{A}_1)F_2(\mathbb{A}_2)] \end{aligned}$$

and two applications of the triangle inequality now complete the job. We have

$$\begin{aligned} |F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{A}_1)F_2(\mathbb{A}_2)| &\leq |F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{B}_1)F_2(\mathbb{B}_2)| + |F_2(\mathbb{B}_2) - F_2(\mathbb{A}_2)| + |F_1(\mathbb{B}_1) - F_1(\mathbb{A}_1)| \\ &\leq |F(\mathbb{A}_1 \times \mathbb{A}_2) - F_1(\mathbb{B}_1)F_2(\mathbb{B}_2)| + F_2(\mathbb{A}_2 \Delta \mathbb{B}_2) + F_1(\mathbb{A}_1 \Delta \mathbb{B}_1) < 8\epsilon. \end{aligned}$$

As ϵ may be chosen arbitrarily small this leaves us with the conclusion that $F(\mathbb{A}_1 \times \mathbb{A}_2) = F_1(\mathbb{A}_1)F_2(\mathbb{A}_2)$ as was to be shown.

While the individual steps of the proof are not difficult, the process is undeniably tedious. The reader who has absorbed the π - λ theorem of Section III.5 may find the following argument more appealing. For $j = 1, 2$, let \mathcal{P}_j be the family of events of the form $\{X_j \leq t\}$ corresponding to the preimages under X_j of the half-closed intervals. As the intersection of two half-closed intervals is another half-closed interval, the families \mathcal{P}_1 and \mathcal{P}_2 form π -classes. If $F(x_1, x_2) = F_1(x_1)F_2(x_2)$ then these classes are independent and, by Theorem III.5.2, it follows that the associated σ -algebras $\sigma(\mathcal{P}_1) = \sigma(X_1)$ and $\sigma(\mathcal{P}_2) = \sigma(X_2)$ are independent. This proof is certainly more compact; but that is because the heavy lifting has already been done by the π - λ theorem. ►

The equivalence of (8.1, 8.1', 8.1'') shows that product measures do indeed exist; as we had anticipated in Section VII.5 any product of two marginal d.f.s yields a two-dimensional d.f. which, in turn, induces a unique product measure on the Borel sets of the plane.

It should not be surprising that functions of independent variables are independent.

THEOREM 2 Suppose X_1 and X_2 are independent random variables and g_1 and g_2 are Baire functions on the real line. Then the random variables $Y_1 = g_1(X_1)$ and $Y_2 = g_2(X_2)$ are also independent.

PROOF: Let G be the two-dimensional d.f. of the random pair (Y_1, Y_2) and G_1 and G_2 the associated marginal d.f.s of Y_1 and Y_2 , respectively. Then

$$\begin{aligned} G(\mathbb{A}_1 \times \mathbb{A}_2) &= P\{Y_1 \in \mathbb{A}_1, Y_2 \in \mathbb{A}_2\} = P\{X_1 \in g_1^{-1}\mathbb{A}_1, X_2 \in g_2^{-1}\mathbb{A}_2\} \\ &= F(g_1^{-1}\mathbb{A}_1 \times g_2^{-1}\mathbb{A}_2) = F_1(g_1^{-1}\mathbb{A}_1)F_2(g_2^{-1}\mathbb{A}_2) = G_1(\mathbb{A}_1)G_2(\mathbb{A}_2), \end{aligned}$$

for any pair of Borel sets \mathbb{A}_1 and \mathbb{A}_2 . ►

The arguments and results carry over with only notational changes to systems of n variables.

DEFINITION 2 Random variables X_1, \dots, X_n defined on the same probability space are *independent* if, and only if,

$$\mathbf{P}\{X_1 \in \mathbb{A}_1, \dots, X_n \in \mathbb{A}_n\} = \mathbf{P}\{X_1 \in \mathbb{A}_1\} \times \dots \times \mathbf{P}\{X_n \in \mathbb{A}_n\}$$

for every choice of Borel sets $\mathbb{A}_1, \dots, \mathbb{A}_n$ of the line.

By identifying selected sets \mathbb{A}_i with the real line it is easy to see that if X_1, \dots, X_n are independent then so are the variables in any subcollection. Indeed, let X_{j_1}, \dots, X_{j_k} be any subcollection. Set $\mathbb{A}_i = \mathbb{R}$ for $i \notin \{j_1, \dots, j_k\}$. Then

$$\mathbf{P}\{X_{j_1} \in \mathbb{A}_{j_1}, \dots, X_{j_k} \in \mathbb{A}_{j_k}\} = \mathbf{P}\{X_{j_1} \in \mathbb{A}_{j_1}\} \times \dots \times \mathbf{P}\{X_{j_k} \in \mathbb{A}_{j_k}\} \quad (8.2)$$

so that X_{j_1}, \dots, X_{j_k} are also independent. In particular, if X_1, \dots, X_n is a system of independent random variables then each pair of random variables X_i, X_j of the collection is independent. The converse, however, is not true as we saw in Examples III.2.7.8.

Suppose the random vector (X_1, \dots, X_n) has d.f. $F(x_1, \dots, x_n)$, the corresponding marginal d.f.s being $F_1(x_1), \dots, F_n(x_n)$. Then (8.2) is equivalent to the statement

$$F(\mathbb{A}_1 \times \dots \times \mathbb{A}_n) = F_1(\mathbb{A}_1) \times \dots \times F_n(\mathbb{A}_n) \quad (\mathbb{A}_1, \mathbb{A}_2, \dots, \mathbb{A}_n \in \mathcal{B}). \quad (8.2')$$

A measure F with this property is said to be a *product measure* and we write $F = F_1 \otimes \dots \otimes F_n$ when we wish to emphasise the product decomposition of the measure. When the marginal measures are all identical, $F_1 = \dots = F_n = \mathfrak{M}$, we say that X_1, \dots, X_n is generated by *independent sampling according to \mathfrak{M}* and write compactly $F = \mathfrak{M}^{\otimes n}$ to mean the product of \mathfrak{M} with itself n times. By restricting the Borel sets to rectangles we obtain a statement corresponding to (8.1''),

$$F(x_1, \dots, x_n) = F_1(x_1) \times \dots \times F_n(x_n). \quad (8.2'')$$

I will note without further comment the n -dimensional counterparts of Theorems 1 and 2; the proofs follow the same pattern as for two dimensions.

THEOREM 1' Random variables X_1, \dots, X_n are independent if, and only if, their n -dimensional d.f. may be factored into a product of marginal d.f.s, $F(x_1, \dots, x_n) = F_1(x_1) \cdots F_n(x_n)$.

THEOREM 2' Suppose X_1, \dots, X_n are independent random variables and g_1, \dots, g_n Baire functions, possibly vector-valued. Then $Y_1 = g_1(X_1), \dots, Y_n = g_n(X_n)$ are also independent.

The final stage of our development is to consider a countable collection of random variables.

DEFINITION 3 A sequence of random variables $\{X_n, n \geq 1\}$ is *independent* if, and only if, for every n , the variables X_1, \dots, X_n are independent.

Our definition is equivalent to the statement that every *finite* subcollection of the countable collection $\{X_n, n \geq 1\}$ is independent. This then is the final step in the development of an abstract notion of independence. The reader should see the similarity in the systematically increasing domain of definition with the corresponding definitions for events in Section III.1. The definitions of this section form the natural generalisation of the abstract concept of independence from sets to random variables or, what is the same thing, to the σ -algebra of sets generated by the random variables. The earlier definitions indeed form a special case of the theory we have developed: if $\{A_n, n \geq 1\}$ is a sequence of events, their independence is equivalent to the statement that the corresponding sequence of indicator random variables $\{1_{A_n}, n \geq 1\}$ is independent.

Now it is certainly intuitive that derived random variables that are functions of non-overlapping collections of independent random variables are themselves independent. The following result, which may be skipped on a first reading, generalises Theorems 2 and 2'. Some notation greases the wheels.

We recall that $\sigma(X)$, the σ -algebra generated by a random variable X , is the smallest σ -algebra with respect to which X is measurable. We broaden the definition to arbitrary families of random variables and write $\sigma(X_\lambda, \lambda \in \Lambda)$ for the smallest σ -algebra with respect to which each of the random variables X_λ in an arbitrary collection is measurable. The concept of independent σ -algebras was introduced in Section III.5.

THEOREM 2'' Suppose $\{X_n, n \geq 1\}$ is a sequence of independent random variables. Let N_1, N_2, \dots denote a countable partition of the natural numbers \mathbb{N} into non-overlapping sets. Then $\{\sigma(X_n, n \in N_k), k \geq 1\}$ is an independent family of σ -algebras. *A fortiori*, if, for each k , Y_k is a Baire function of $(X_n, n \in N_k)$, then $\{Y_k, k \geq 1\}$ is an independent sequence.

PROOF: The demonstration uses the $\pi-\lambda$ theorem of Section III.5 in an essential way. Let \mathcal{P}_k be the π -class of finite intersections of sets of the form $\{X_n \leq t\}$ with $n \in N_k$ and $t \in \mathbb{R}$. As independence is defined in terms of finite collections it follows immediately that $\{\mathcal{P}_k, k \geq 1\}$ is an independent sequence of π -classes. Indeed, suppose \mathbb{K} is any finite collection of positive numbers and, for each $k \in \mathbb{K}$, let A_k be an element of \mathcal{P}_k . Then A_k is a finite intersection of sets of the form $\{X_n \leq t_n\}$ where n varies over some finite subset J_k of N_k . The variables in the subcollection $\{X_n, n \in J_k, k \in \mathbb{K}\}$ are certainly independent and hence

$$P\left(\bigcap_{k \in \mathbb{K}} A_k\right) = P\left(\bigcap_{k \in \mathbb{K}} \bigcap_{n \in J_k} \{X_n \leq t_n\}\right) = \prod_{k \in \mathbb{K}} \prod_{n \in J_k} P\{X_n \leq t_n\} = \prod_{k \in \mathbb{K}} P(A_k).$$

And so $\{\mathcal{P}_k, k \geq 1\}$ is independent. As $\sigma(\mathcal{P}_k) = \sigma(X_n, n \in N_k)$, the claimed result follows directly as a consequence of Theorem III.5.2. ▶

It is occasionally useful in applications to remark that the definitions and results carry through barring only slight notational changes to independent vector-valued random variables X_1, \dots, X_n, \dots , not necessarily in the same number of dimensions, where we now naturally interpret F_k as the (multi-dimensional) d.f. of X_k , with g_k a (possibly vector-valued) Baire function (in the appropriate number of dimensions). All that is requisite is to strike out the odd references to intervals and the real line and replace them by rectangles and the appropriate finite-dimensional Euclidean space.



9 Do independent variables exist?

While we have seen how to construct product measures in a finite number of dimensions, it is not at all clear how to proceed when we move to an infinite number of dimensions. Do independent random variables exist? More specifically, given a probability space (Ω, \mathcal{F}, P) , can we construct a sequence of independent random variables of given marginal distributions on this space? The answer is "Yes", but only by *imbedding* a product structure in the space. The intuitive content of the theory suggests the way.

Our common experience suggests that we may consider a sequence of coin tosses as "independent" and it is natural to seek to extend this intuition to a *gedanken* experiment involving an unending sequence of tosses. In Chapter V we saw how this could be formally accomplished by associating the sample points of such an experiment with points in a continuum. Reversing the process, suppose Z is a random variable distributed uniformly in the unit interval $[0, 1]$. Eschewing unnecessary generality, we may consider the unit interval $[0, 1]$ as sample space equipped with Lebesgue measure λ on the Borel sets of $[0, 1]$: $P\{Z \leq t\} = \lambda[0, t] = t$ for $0 \leq t < 1$. We may, as before, express Z in its dyadic representation $Z = Z_1 2^{-1} + Z_2 2^{-2} + Z_3 2^{-3} + \dots$ where $Z_k \in \{0, 1\}$ for each k and, for definiteness, when there are two representations for Z we select the one with an infinity of zeros.

LEMMA 1 *The random variables $\{Z_k, k \geq 1\}$ form a sequence of Bernoulli trials corresponding to repeated tosses of a fair coin.*

This is simply a restatement of the major conclusion of Section V.3 that the binary digits are independent. The sequence $\{Z_k, k \geq 1\}$ is hence independent and we have managed to imbed a product measure in $[0, 1]$. This admittedly looks like a very special case but, as the following development shows, is actually of great generality.

The idea is to extract independent subsequences from the sequence $\{Z_k, k \geq 1\}$. Let $(i, j) \mapsto k(i, j)$ be any one-to-one map taking each pair (i, j) of natural numbers into a unique natural number $k = k(i, j)$. (Cantor's diagonal method will do; see Figure XI.1.) We set $Z_{ij} = Z_k$ whence the double sequence $\{Z_{ij}, i \geq 1, j \geq 1\}$ constitutes a renumbering of the original sequence into independent binary subsequences

$$\begin{array}{ccccccc} Z_{11} & Z_{12} & Z_{13} & \dots & Z_{1n} & \dots \\ Z_{21} & Z_{22} & Z_{23} & \dots & Z_{2n} & \dots \\ \dots & \dots & \dots & & \dots & \dots \\ Z_{n1} & Z_{n2} & Z_{n3} & \dots & Z_{nn} & \dots \\ \dots & \dots & \dots & & \dots & \dots \end{array}$$

For each n , the subsequence $\{Z_{nj}, j \geq 1\}$ constituting the n th row of the array represents the binary digits in the dyadic expansion of the variable $U_n = Z_{n1}2^{-1} + Z_{n2}2^{-2} + Z_{n3}2^{-3} + \dots$. It follows that U_n is uniformly distributed in the unit interval $[0, 1]$. As each U_n is determined by a non-overlapping set of the Z_{ij} it is tempting to conclude that the variables U_1, U_2, \dots are indeed independent. And indeed they are though we will have to negotiate the technical issue that each of the variables U_n is determined by a denumerably infinite collection of Bernoulli variables.

LEMMA 2 *The random variables $U_1, U_2, \dots, U_n, \dots$ are independent, each distributed uniformly in the unit interval $[0, 1]$.*

PROOF: Let $U'_1 = \sum_{k=1}^m Z_{1k}2^{-k}$ be the truncation of the dyadic expansion of U_1 to m terms. Writing $U''_1 = U_1 - U'_1$, the uniform convergence of the dyadic expansion shows that $0 \leq U''_1 < \sum_{k=m+1}^{\infty} 2^{-k} = 2^{-m}$ and it follows that

$$P\{U'_1 \leq t_1 - 2^{-m}\} \leq P\{U_1 \leq t_1\} \leq P\{U'_1 \leq t_1 + 2^{-m}\}.$$

The random variable U'_1 is discrete and places equal mass on each of the 2^m points $0/2^m, 1/2^m, \dots, (2^m - 1)/2^m$. Accordingly, for each $0 \leq t'_1 < 1$, we have

$$P\{U'_1 \leq t'_1\} = \frac{\lfloor t'_1 2^m \rfloor}{2^m} = t'_1 - \frac{\delta_1}{2^m}$$

where $0 \leq \delta_1 = \delta_1(t'_1, m) < 1$. For any $0 < t_1 < 1$ we may select m so large that $2^{-m} < \min\{t_1, 1 - t_1\}$ whence $t_1 \pm 2^{-m}$ is confined to the unit interval. Identifying t'_1 with $t_1 - 2^{-m}$ and $t_1 + 2^{-m}$ in turn, we see that

$$t_1 - 2^{-m+1} \leq P\{U_1 \leq t_1\} \leq t_1 + 2^{-m+1},$$

and by taking the limit as m tends to infinity we obtain $P\{U_1 \leq t_1\} = t_1$. The same argument may be applied to any of the variables U_n and it follows that the variables U_n are all uniformly distributed in the unit interval.

To complete the proof we have to show that U_1, \dots, U_n are independent for any n . We may appeal to Theorem 8.2'' directly but a direct verification has its charms. It will suffice to consider the case $n = 2$. With the selection of the nonce variables $U'_2 = \sum_{k=1}^m Z_{2k}2^{-k}$ and $U''_2 = U_2 - U'_2$ paralleling that of the corresponding variables U'_1 and U''_1 , we have

$$\begin{aligned} P\{U'_1 \leq t_1 - 2^{-m}, U'_2 \leq t_2 - 2^{-m}\} &\leq P\{U_1 \leq t_1, U_2 \leq t_2\} \\ &\leq P\{U'_1 \leq t_1 + 2^{-m}, U'_2 \leq t_2 + 2^{-m}\}. \end{aligned}$$

The functions $U'_1 = U'_1(Z_{11}, \dots, Z_{1m})$ and $U'_2 = U'_2(Z_{21}, \dots, Z_{2m})$ are determined by non-overlapping finite sets of independent variables. It follows hence that U'_1 and U'_2 are independent, their common distribution placing equal mass on the values $0/2^m, 1/2^m, \dots, (2^m - 1)/2^m$. Thus, for any $0 < t'_1, t'_2 < 1$, we have

$$\begin{aligned} P\{U'_1 \leq t'_1, U'_2 \leq t'_2\} &= P\{U'_1 \leq t'_1\} P\{U'_2 \leq t'_2\} \\ &= \frac{\lfloor t'_1 2^m \rfloor}{2^m} \cdot \frac{\lfloor t'_2 2^m \rfloor}{2^m} = \left(t'_1 - \frac{\delta_1}{2^m} \right) \left(t'_2 - \frac{\delta_2}{2^m} \right) \end{aligned}$$

where $0 \leq \delta_1, \delta_2 < 1$. For any $0 < t_1, t_2 < 1$, we may now select m so large that $2^{-m} < \min\{t_1, t_2, 1-t_1, 1-t_2\}$ and argue as in the case of one variable to obtain

$$(t_1 - 2^{-m+1})(t_2 - 2^{-m+1}) \leq \mathbf{P}\{U_1 \leq t_1, U_2 \leq t_2\} \leq (t_1 + 2^{-m+1})(t_2 + 2^{-m+1}).$$

Proceeding to the limit as m tends to infinity now establishes that U_1 and U_2 are indeed independent. The argument for arbitrary n proceeds along entirely similar lines. ►

This is progress. We have now managed to imbed a sequence $\{U_n, n \geq 1\}$ of independent, uniformly distributed random variables in the original sample space consisting of the unit interval $[0, 1]$ equipped with Lebesgue measure λ on the Borel sets. We only need now to morph the constructed variables to obtain a sequence of independent variables with given marginal distributions.

Let U represent a generic random variable distributed uniformly in the unit interval $[0, 1]$. Let F be any d.f. The salient point of the construction is the map $F^\dagger: u \mapsto x_u$ which, to each point $0 < u < 1$, associates the value $x_u = \inf\{x : F(x) \geq u\}$.

LEMMA 3 *The random variable $X = F^\dagger(U)$ has d.f. F .*

PROOF: As F is a d.f., hence continuous from the right, $F(x) \geq u$ if $x \geq x_u$ and $F(x) < u$ if $x < x_u$. It follows that $X \leq x$, or, equivalently, $x_u \leq x$, if, and only if, $U \leq F(x)$. ►

All is now in place. Let F_1, F_2, \dots be any sequence of d.f.s. For each $n \geq 1$, we set $X_n = F_n^\dagger(U_n)$. Then X_n has d.f. F_n and, as functions of independent variables are independent, we have constructed a sequence of independent random variables of specified marginal distribution.

THE FUNDAMENTAL EXISTENCE THEOREM OF PRODUCT MEASURE *Let $\{F_n, n \geq 1\}$ be any sequence of d.f.s. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a sequence of independent random variables $\{X_n, n \geq 1\}$ on this space such that, for each n , X_n has d.f. F_n .*

Our construction of independent random variables may be traced ultimately back, $Z \mapsto \{Z_k, k \geq 1\} \mapsto \{Z_{ij}, i \geq 1, j \geq 1\} \mapsto \{U_n, n \geq 1\} \mapsto \{X_n, n \geq 1\}$, so that $X_n = X_n(Z)$ ($n \geq 1$) constitutes an imbedding of an independent sequence in the original space. The procedure is certainly indirect and the reader may wonder if it is possible to work directly with the *infinite product space* \mathbb{R}^∞ consisting of denumerably infinite sequences $(x_1, x_2, \dots, x_n, \dots)$ of real numbers. This will require (a) the identification of a suitable σ -algebra \mathcal{F} of subsets of \mathbb{R}^∞ whose projections, for each n , onto the first n coordinates yield the Borel sets of \mathbb{R}^n , and (b) the construction of a probability measure F on \mathcal{F} whose projection, for each n , onto the first n coordinates yields the product measure $F_1 \times \dots \times F_n$. I will not go into the details of the construction here but this is indeed possible as Kolmogorov showed in his seminal work of 1933. This first existence theorem is of fundamental importance in the theory of stochastic processes.



10 Remote events are either certain or impossible

In many situations involving independent random variables we are interested in events that are determined by the random variables “out at infinity”. Here is a specific illustration. Suppose X_1, X_2, \dots is a sequence of independent random variables. Let A be

the set of sample points ω for which $n^{-1}(X_1(\omega) + \dots + X_n(\omega)) \rightarrow 0$ as $n \rightarrow \infty$. For any fixed m , it is clear that $n^{-1}(X_1(\omega) + \dots + X_m(\omega)) \rightarrow 0$ as $n \rightarrow \infty$. And hence, $\lim_n n^{-1} \sum_{j=1}^n X_j(\omega) = 0$ if, and only if, $\lim_n n^{-1} \sum_{j=m}^n X_j(\omega) = 0$. Thus, we may write A by a countable number of operations via

$$A = \bigcap_{k \geq 1} \bigcup_{v \geq m} \bigcap_{n \geq v} \left\{ \omega : \left| \frac{1}{n} \sum_{j=m}^n X_j(\omega) \right| < \frac{1}{k} \right\}. \quad (10.1)$$

The event inside the innermost intersection is determined only by the random variables X_m, X_{m+1}, \dots and hence so is A . As m may be chosen arbitrarily large, A is effectively determined by the random variables arbitrarily far out in the sequence. This suggests that we consider the σ -algebras generated by the tail.

Introduce the “tail sequence” of σ -algebras $\{\mathcal{T}_n, n \geq 1\}$ where, for each n , $\mathcal{T}_n = \sigma(X_n, X_{n+1}, \dots)$ is the minimal σ -algebra with respect to which each of X_n, X_{n+1}, \dots is measurable. It is clear that $\{\mathcal{T}_n, n \geq 1\}$ is a decreasing sequence with limit a σ -algebra $\mathcal{T} = \bigcap_n \mathcal{T}_n$ contained in each of the σ -algebras \mathcal{T}_n . We call \mathcal{T} the *tail σ -algebra* of the sequence $\{X_n, n \geq 1\}$ and refer to events in \mathcal{T} as *tail events* or, more vividly, *remote events*. Kolmogorov discovered that remote events have a very simple characterisation.

KOLMOGOROV’S ZERO–ONE LAW *Every remote event has probability either zero or one only.*

PROOF: The proof of the theorem uses independence in a fundamental way. Suppose A is an event in the tail σ -algebra \mathcal{T} . Then $A \in \sigma(X_{n+1}, X_{n+2}, \dots)$ for each n and, *a fortiori*, $A \in \sigma(X_1, X_2, \dots)$. Now fix any $n \geq 1$. By Theorem 8.2”, the σ -algebras $\sigma(X_1, \dots, X_n)$ and $\sigma(X_{n+1}, X_{n+2}, \dots)$ are independent. Suppose A' is an event in $\sigma(X_1, \dots, X_n)$. As A is in $\sigma(X_{n+1}, X_{n+2}, \dots)$, it follows that A and A' are independent events [or, what is the same thing, the indicator 1_A is independent of (X_1, \dots, X_n)]. As this is true for every n , this means that A is independent of the events in $\sigma(X_1, X_2, \dots)$. Thus, A is independent of itself, or, $P(A) = P(A \cap A) = P(A) \cdot P(A)$. But this means $P(A) = 0$ or $P(A) = 1$. ►

Remote events crop up naturally in a variety of circumstances as the following examples attest. In these situations it is comforting to know that event probabilities can only be zero or one. But it can be surprisingly hard to figure out which of the two possibilities actually holds sway.

EXAMPLES: 1) *Convergent series.* By an argument similar to (10.1), the set of sample points on which the sequence of partial sums $X_1 + \dots + X_n$ converges is a remote event, hence has probability either zero or one. Which one depends on the nature of the underlying distribution. We shall return to the problem armed with a new tool in Section XVI.10.

2) *Run lengths.* Suppose X_1, X_2, \dots is a sequence of symmetric Bernoulli trials. In the notation of Section VIII.8, for each n , let R_n represent the length of the run of failures beginning at n . Let $\{r_n, n \geq 1\}$ be any given positive sequence. Then the event $\{R_n \geq r_n \text{ i.o.}\}$ only depends upon the remote tails. Indeed, for any m , we have

$$\{R_n \geq r_n \text{ i.o.}\} = \bigcap_{n \geq m} \bigcup_{k \geq n} \{R_k \geq r_k\}$$

and as the occurrence of the event $\{R_k \geq r_k\}$ is completely determined by the random variables $X_k, X_{k+1}, \dots, X_{k+r_k-1}$, it follows that $\{R_n \geq r_n \text{ i.o.}\}$ lies in $\sigma(X_m, X_{m+1}, \dots)$. As this is true for every m , it follows that $\{R_n \geq r_n \text{ i.o.}\}$ lies in the tail σ -algebra of the sequence $\{X_n, n \geq 1\}$. Thus, $\{R_n \geq r_n \text{ i.o.}\}$ is a remote event and can hence take probabilities zero or one only. Which of these possibilities is true depends upon the sequence $\{r_n\}$.

Say that $\{r_n\}$ is an *outer envelope* or an *inner envelope* depending on whether the probability of the remote event $\{R_n \geq r_n \text{ i.o.}\}$ is zero or one, respectively. Then (VIII.8.1) and (VIII.8.2) show that $r_n = \alpha \log_2 n$ is an outer envelope if $\alpha > 1$ and an inner envelope if $\alpha \leq 1$. We can obtain a slightly more refined picture.

Suppose $r_n = \log_2 n + \alpha \log_2 \log_2 n$ for $n \geq 2$. Here α is a constant. If $\alpha > 1$ then the sum $\sum_n 2^{-r_n} = \sum_n 1/(n \log_2(n)^\alpha)$ converges and the argument leading up to (VIII.8.1) shows that $\alpha > 1$ results in an outer envelope. On the other hand, $2^{-r_n}/r_n$ is of the order of $1/(n \log_2(n)^{1+\alpha})$ and if $\alpha \leq 0$ the sum $\sum_n 2^{-r_n}/r_n$ diverges; by the analysis leading to (VIII.8.2), we then have an inner envelope for $\alpha \leq 0$. The cases $0 < \alpha \leq 1$ will yield either an outer or an inner envelope but this analysis is not subtle enough to discover which. ▶

11 Problems

1. The reader will have realised that the proof given in the text that the points of jump of a d.f. are countable does not rely upon the boundedness of the d.f. and indeed carries over verbatim to increasing functions, not necessarily bounded. Here is another proof. Let f be an increasing real-valued function of a real variable. If f is bounded, that is, there are A and B such that $A \leq f(x) \leq B$ for all x , show that, for every $\epsilon > 0$, the number of jumps of size exceeding ϵ is no more than $(B - A)/\epsilon$. Hence argue that the number of jumps of f is countable, first for bounded f , and then in general.

2. *Inverse mapping.* Suppose X is any real-valued (or extended real-valued) map on Ω . Suppose \mathbb{A} is any subset of the real line. Show that $X^{-1}(\mathbb{A}^c) = (X^{-1}(\mathbb{A}))^c$.

3. *Continuation.* Suppose $\{\mathbb{A}_\lambda\}$ is any collection of subsets of the real line where λ ranges over an arbitrary index set, not necessarily countable. Show that $X^{-1}(\bigcup_\lambda \mathbb{A}_\lambda) = \bigcup_\lambda X^{-1}(\mathbb{A}_\lambda)$ and $X^{-1}(\bigcap_\lambda \mathbb{A}_\lambda) = \bigcap_\lambda X^{-1}(\mathbb{A}_\lambda)$.

4. A plausible definition of a discrete d.f. is that “it is constant between jumps”. Why is this definition insufficient?

5. *Balls and urns.* Suppose r balls are randomly distributed in n urns. Let X be the number of empty urns. With n fixed, write $p_r(m) = P[X = m]$ for the distribution of X parametrised by r . Show that the recurrence

$$p_{r+1}(m) = p_r(m) \frac{n-m}{n} + p_r(m+1) \frac{m+1}{n} \quad (11.1)$$

holds for all m and r and solve it.

6. *Maximum, minimum.* Introduce the short-hand notation $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. Suppose X and Y are independent random variables with distribution functions $F(x)$ and $G(y)$, respectively. Determine the distributions of $X \vee Y$ and $X \wedge Y$.

7. *Continuation.* Suppose Z is a Bernoulli variable with success probability p that is independent of X and Y . Determine the distribution functions of (a) $R = ZX + (1-Z)Y$, (b) $S = ZX + (1-Z)(X \wedge Y)$, and (c) $T = ZX + (1-Z)(X \vee Y)$.

8. Let F be a d.f. with points of jump $\{a_j\}$. Show that, for every x , as $\epsilon \downarrow 0$, the sum $\sum_{j: x-\epsilon < a_j < x} [F(a_j) - F(a_j-)]$ converges to zero. What if the summation is extended to $x - \epsilon < a_j \leq x$ instead?

9. *Continuation.* Let $b_j = F(a_j) - F(a_j-)$ be the size of the jump of F at a_j and let $F_1(x) = \sum_j b_j H_0(x - a_j)$ be the discrete part of F . Show that the function $F_2(x) = F(x) - F_1(x)$ is continuous. Conclude that every d.f. may be written as a convex combination $F = \alpha F_d + (1 - \alpha) F_c$ where $0 \leq \alpha \leq 1$, F_d is a discrete d.f., and F_c is a continuous d.f.

10. *Continuation, Lebesgue's Decomposition Theorem.* A theorem of Lebesgue's asserts that every d.f. F has a positive derivative F' a.e. (with respect to Lebesgue measure). Using this show that every d.f. may be written as a convex combination $F = \alpha F_d + \beta F_{ac} + \gamma F_{sc}$ where $\alpha, \beta, \gamma \geq 0$, $\alpha + \beta + \gamma = 1$, F_d is discrete, F_{ac} is absolutely continuous, and F_{sc} is singular continuous, such a decomposition being unique.

11. Within the unit square put $F(x, y) = x$ if $x \leq y$ and $F(x, y) = y$ if $x > y$. Show that F is a distribution concentrated at the bisector (hence singular).

12. *The Cantor distribution.* Any point x in the unit interval may be expressed in a ternary expansion $x = \sum_{n=1}^{\infty} c_n 3^{-n}$ where $c_n \in \{0, 1, 2\}$ for each n . Argue inductively that, for each n , we may interpret the ternary value c_n as determining whether x lies to the left of, in, or to the right of a middle-thirds open interval of length $3^{-(n-1)}$. With this convention, argue hence that for x to be in the open set \mathbb{U} of Example XI.3.3, it is necessary and sufficient that at least one ternary digit c_n take the value 1; or, equivalently, any point x in the ternary Cantor set \mathbb{C} has a ternary expansion with each digit c_n taking values 0 or 2 only. Prove that for any $x \in \mathbb{C}$ we have $F(x) = \sum_{n=1}^{\infty} c_n 2^{-n-1}$ where F is the associated ternary Cantor d.f.

13. *Continuation.* For each x in $[0, 1]$ show that the ternary Cantor d.f. satisfies $F\left(\frac{x}{3}\right) = \frac{1}{2} F(x)$ and $F\left(\frac{2}{3} + \frac{x}{3}\right) = \frac{1}{2} (1 + F(x))$.

14. Let $\{X_k, k \geq 1\}$ be a sequence of Bernoulli trials, each X_k taking values 0 and 1 with probability one-half apiece. Let $T_0 = \sum_{k=1}^{\infty} X_{3k} 2^{-3k}$, $T_1 = \sum_{k=1}^{\infty} X_{3k-1} 2^{-3k+1}$, and $T_2 = \sum_{k=1}^{\infty} X_{3k-2} 2^{-3k+2}$. Determine explicitly the nature of the d.f.s of T_0 , T_1 , and T_2 . What is the d.f. of $S = T_0 + T_1 + T_2$?

15. *Random directions in the plane.* Suppose \mathbf{X} and \mathbf{Y} are independent unit vectors with random directions in \mathbb{R}^2 (endpoints uniformly distributed on the unit circle). Show that the length L of the resultant $\mathbf{S} = \mathbf{X} + \mathbf{Y}$ has density $2/(\pi\sqrt{4-t^2})$ concentrated on $(0, 2)$. [Hint: Use the law of cosines.]

16. *Random directions in \mathbb{R}^3 .* Let L be the length of the resultant of two independent unit vectors with random directions in \mathbb{R}^3 (endpoints uniformly distributed on the unit sphere). Show that $P\{L \leq t\} = t^2/2$ for $0 < t < 2$.

17. *Fair coins from bent coins.* Repeated tosses of a bent coin give rise to a Bernoulli sequence X_1, X_2, \dots with success probability p . Devise a procedure that constructs a sequence of fair coin flips, i.e., a sequence of Bernoulli trials Z_1, Z_2, \dots with success probability $1/2$, from the given bent sequence.

18. *The uniform density.* Let X be a random variable with a continuous d.f. $F(x)$. Form the random variable $Y = F(X)$. Show that Y is uniformly distributed in the unit interval. [Hint: Build intuition by considering a strictly increasing d.f. first before generalising your argument.]

19. *Empirical distributions.* Let X_1, \dots, X_n be independent random variables with a common continuous distribution F . The associated empirical d.f. F_n is the discrete d.f. with jumps of size $1/n$ at the points X_1, \dots, X_n . For each x , the random variable $F_n(x) = \sum_{j=1}^n 1(X_j \leq x)$ is close to $F(x)$ when n is large by the law of large numbers. The maximum discrepancy $D_n = \sup_x |F_n(x) - F(x)|$ is hence of interest to statisticians. Using the previous problem argue that the distribution of D_n does not depend on the choice of F ; in particular, we may just as well assume that F is uniform.

20. *Continuation, symmetrisation.* Suppose X_1, \dots, X_n and $X_1^\#, \dots, X_n^\#$ is an independent double sample drawn from a common continuous distribution F . Let F_n and $F_n^\#$, respectively, denote the empirical distributions of the two samples. The maximum discrepancy between the two empirical distributions is defined by $D_{n,n} = \sup_x |F_n(x) - F_n^\#(x)|$. Show that the distribution of $D_{n,n}$ does not depend on the choice of F .

21. *The Box–Muller construction.*² Suppose U and V are independent random variables, each distributed uniformly in the unit interval. Form the random variables $R = \sqrt{-2 \log(1-U)}$ and $\Theta = 2\pi V$ and determine their distributions. Now let $X = R \cos \Theta$ and $Y = R \sin \Theta$. What is their joint distribution?

22. *Ladder indices.* The random variable N is defined as the unique index such that $X_1 \geq X_2 \geq \dots \geq X_{N-1} < X_N$. If the X_i are independent and have a common continuous distribution F , determine the distribution $p(n) := P\{N = n\}$ and its expected value $E(N)$.

23. *Continuation.* Suppose explicitly that the d.f. F of the previous problem is the uniform distribution in the unit interval $0 < x < 1$. Prove that

$$P\{X_1 \leq x, N = n\} = \frac{x^{n-1}}{(n-1)!} - \frac{x^n}{n!}.$$

24. *Continuation, von Neumann's construction.*³ Under the conditions of the previous problems, define Y as follows: a “trial” is a sequence X_1, \dots, X_N ; it is a “failure” if N is odd, a “success” if N is even. We repeat independent trials as long as necessary to produce a “success”. Let Y equal the number of failures plus the first variable in the successful trial. Determine $P\{Y < x\}$. Looking end-to-end, comment on what this procedure has wrought.

25. *Run lengths.* Refine the analysis of Example 10.2 and show that $r_n = \log_2 n + \log_2 \log_2 n + \alpha \log_2 \log_2 \log_2 n$ is an outer envelope if $\alpha > 1$ and that $r_n = \log_2 n + \log_2 \log_2 \log_2 n$ is an inner envelope.

26. Let A_1, A_2, \dots be independent events. Show that the set of sample points ω on which $\frac{1}{n} \sum_{k=1}^n 1_{A_{n_k}}(\omega) \rightarrow x$ has probability zero or one.

²G. E. P. Box and M. E. Muller, “A note on the generation of random normal deviates”, *Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 610–611, 1958.

³J. von Neumann, *National Bureau of Standards, Appl. Math. Series*, no. 12, pp. 36–38, 1951.

XIII

Great Expectations

The distribution function of a random variable tells us all there is to know, statistically speaking, about it. In many settings of theoretical and practical interest, however, one is satisfied with cruder information about the variable. (And if the distribution in question is unknown or only partly known, one may have to resort to these cruder information measures in any case to make a virtue out of necessity.) While a variety of measures can be developed along these lines, principal among them are measures of central tendency in a random variable.

C 1–5, 8
A 6, 7

1 Measures of central tendency

Efforts to simply characterise random variables are represented by questions of the following stripe. What is the “typical” value of a random variable? Where does a random variable “spend most of its time”? What would be a good *a priori* guess for the value of a random variable? As we shall see, the attempt to mathematically capture the intuitive content of these questions leads to a particularly fecund theory.

The simplest descriptors of central tendency in a random variable include the *mean*, the *median*, and the *mode*. The following example is illustrative.

EXAMPLE: *A waiting time problem.* Consider a succession of independent tosses of a bent coin which turns up “Heads” with probability p . The tosses are continued until a “Head” is first observed. Let the random variable X denote the waiting time till success, i.e., the number of failures before the occurrence of the “Head”. Then X conforms to a geometric distribution and for each positive integer value $k \geq 0$, $w_k := P\{X = k\} = q^k p$ where, as usual, we write $q = 1 - p$. What can we say about the “typical” values of X ?

A natural candidate is the *mean* of the distribution which we recall is given by the weighted sum $\mu = 0w_0 + 1w_1 + 2w_2 + 3w_3 + \dots$. This is usually the notion referred to when one speaks of an “average” or a “centre of mass”. For the geometric distribution, as we’ve already seen, $\mu = \sum_{k=0}^{\infty} kq^k p = q/p$.

Thus, if the coin is fair, on average one would expect to see one “Tail” before encountering the first “Head”.

As an alternative measure of “typicality” one may choose to consider the “midway” point M characterised by the twin inequalities $w_0 + w_1 + \dots + w_{M-1} < \frac{1}{2}$ and $w_0 + w_1 + \dots + w_M \geq \frac{1}{2}$, where it is about as likely that the tosses will continue beyond that point as it is that the tosses terminate before it. This is called the *median* of the distribution. In our example,

$$\frac{1}{2} > \sum_{k=0}^{M-1} w_k = p \sum_{k=0}^{M-1} q^k = \frac{p(1-q^M)}{1-q} = 1 - q^M,$$

whereas, on the other hand,

$$\frac{1}{2} \leq \sum_{k=0}^M w_k = \frac{p(1-q^{M+1})}{1-q} = 1 - q^{M+1}.$$

The median of the geometric distribution is hence the unique integer M satisfying

$$\frac{-\log 2}{\log(1-p)} - 1 \leq M < \frac{-\log 2}{\log(1-p)}.$$

If p is small then the median of the waiting time to achieve a “Head” is approximately $\frac{1}{p} \log 2$. For a fair coin, the median is $M = 0$, as is easy to see because $w_0 = q^0 p = 1/2$ in the case when the coin is fair.

Yet another notion of “typical” arises when we identify X with its most probable value $m = \arg \max_k w_k$. This is called the *mode*. For our waiting time problem, it is clear that the probabilities $w_k = q^k p$ decrease monotonically with k whence $m = 0$ for all values of p . ▶

It is worth comparing the very different flavours of these three notions of “typicality”. Which is best? Each has utility but, while no single measure of central tendency is completely satisfactory in all applications, the mean, also called the *expectation*, is by far the most important and the most prevalent in theory and application.

The idea of expectation that we’ve encountered hitherto in discrete and continuous settings does well enough for many purposes but suffers from an annoying discrepancy in notation—a sum weighted by a probability distribution in the one case and a (Riemann) integral weighted by a density in the other. This is exacerbated in situations where variables exhibit a mixture of discrete and continuous attributes; and in situations where the variables have a singular continuous distribution, the notion, as developed thus far, fails completely. In this chapter we will see a general unified framework and notation that considerably simplifies and demystifies the concept and helps identify key features which otherwise may be overlooked as accidental. As we shall see, we can write

the expectation, when it exists, of an arbitrary random variable X with distribution function $F(x)$ in the compact integral notation

$$E(X) = \int_{-\infty}^{\infty} x dF(x). \quad (1.1)$$

Of course, if X is discrete and takes values in the set $\{x_k, k \geq 1\}$ with corresponding probabilities $\{p_k, k \geq 1\}$ then the expression above reduces to the familiar form $E(X) = \sum_{k \geq 1} x_k p_k$; likewise, if X is continuous with density $f(x)$ then the expression reduces, as it must, to $E(X) = \int_{-\infty}^{\infty} xf(x) dx$.

The rest of this chapter deals with the long-promised development of a more supple and general theory of integration and, in particular, an understanding of what an expression like (1.1) really means. The key properties that emerge are that expectation is additive, $E(X + Y) = E(X) + E(Y)$, homogeneous, $E(cX) = cE(X)$, and monotone, $E(X) \leq E(Y)$ if $X \leq Y$. As we shall see, one of the ancillary benefits that we shall derive from a more general vantage point is that arguments involving passages to a limit are considerably simplified; for many limiting arguments one can be satisfied with a vague “goes through generally” which appeals implicitly to a powerful and general theory of integration put forward by Henri Lebesgue in 1902. For many applications the following slogan will suffice.

SLOGAN *The Lebesgue theory makes transparent passages to the limit.*

In this chapter we will flesh out the key components of Lebesgue’s theory. The reader who is a little impatient with this leisurely programme and is willing to accept the slogan at face value can jump ahead to sample the applications in the succeeding chapters in almost any order.

2 Simple expectations

The situation is most transparent and intuitive in the case of simple random variables and we consider these first. We recall that a random variable $X = X(\omega)$ is *simple* if it takes values only in a finite range, say, $\{x_1, \dots, x_n\}$. We suppose that X is defined on an arbitrary abstract probability space Ω equipped with a probability measure P . Writing $A_k = \{\omega : X(\omega) = x_k\}$ for the subset of sample points (event) on which X takes value x_k , we observe that the events A_1, \dots, A_n are disjoint and cover the entire sample space, in other words the events $\{A_1, \dots, A_n\}$ form a finite partition of the sample space. We can then write $X(\omega)$ in its so-called *standard representation*

$$X(\omega) = \sum_{k=1}^n x_k 1_{A_k}(\omega). \quad (2.1)$$

If we write $p_k = P(A_k)$ then the *expectation of the simple random variable X is denoted $E(X)$ and defined by $E(X) = \sum_{k=1}^n p_k x_k$* . In words, the expectation of X is the centre of (probability) mass of X. This, of course, coincides with our earlier naïve definition. Degenerate random variables $X(\omega) = x 1_{\Omega}(\omega) = x$ which take a constant value x provide trivial examples of simple random variables—as is natural, even inevitable, $E(X) = x$. The first non-trivial example, while still easy, is yet illustrative.

EXAMPLE 1) *The flip of a coin, indicators.* If X is a Bernoulli random variable with success probability p, that is to say, $P\{X = 1\} = p$ and $P\{X = 0\} = 1 - p$, then X models the flip of a coin and it is trite to see that $E(X) = p$. Now consider an arbitrary event A in an abstract probability space. Then the associated indicator $1_A(\omega)$ which takes value 1 if $\omega \in A$ and value 0 if $\omega \notin A$ is a Bernoulli random variable with success probability $p = P\{1_A = 1\} = P(A)$ and it follows that $E(1_A) = P(A)$. In other words, *event probabilities may be identified with expectations of the corresponding indicators.* ►

Thus, any probability may be written as an expectation. This turns out to be an extremely potent observation as a route is now suggested to *estimating probabilities through expectations*. The reader would do well to keep this in mind for the nonce; we will fully exploit this idea in Chapters XVII and XVIII.

EXAMPLE 2) *Again, the binomial.* The binomial distribution provides another natural example of a simple random variable. Suppose we take as sample space the set of binary n-tuples $(x_1, \dots, x_n) \in \{0, 1\}^n$ equipped with the uniform measure, $P\{(x_1, \dots, x_n)\} = 2^{-n}$ for each $(x_1, \dots, x_n) \in \{0, 1\}^n$. As k ranges from 0 to n the events $A_k = \{(x_1, \dots, x_n) : x_1 + \dots + x_n = k\}$ partition the entire sample space and the simple random variable $S_n(x_1, \dots, x_n) = \sum_{k=0}^n k 1_{A_k}(x_1, \dots, x_n)$ is equipped with the binomial distribution $b_n(k; p) = \binom{n}{k} p^k q^{n-k}$. As seen by elementary calculations in Section VIII.2, $E(S_n) = \sum_k k b_n(k; p) = np$. ►

Our focus now is not on computing expected values for given distributions *per se* but on deducing general properties of the operation. We begin with the observation that it is not necessary to keep to the standard representation of a simple random variable.

We recall that a sequence $\{B_j\}$ of subsets of sample points is called a *measurable partition* of the sample space if (i) each B_j is an event (hence measurable), (ii) the sets are mutually disjoint, $B_j \cap B_k = \emptyset$ if $j \neq k$, and (iii) the sets cover the sample space, $\bigcup_j B_j = \Omega$.

Let $\{A_k\}$ and $\{B_j\}$ be measurable partitions. We say that $\{B_j\}$ is a *refinement* of $\{A_k\}$ if for every B_j there is some A_k containing it. Intuitively, a partition cuts up the space into disjoint sets and a refinement cuts up the sets in a partition into smaller sets. Figure 1 illustrates the idea; the refinement of the original partition now includes sets (dotted lines) contained within the sets of the original partition.



Figure 1: Refinement of a partition.

Let X be a simple random variable taking values in the range $\{x_1, \dots, x_n\}$ with standard representation (2.1). It is clear that the events $A_k = \{\omega : X(\omega) = x_k\}$ form a finite measurable partition of the sample space. Suppose $\{B_j, 1 \leq j \leq m\}$ is any measurable partition that is a refinement of $\{A_k\}$. Then to each B_j there exists a unique $A_k = A_{k(j)}$ such that $B_j \subseteq A_{k(j)}$, whence $X(\omega) = x_{k(j)}$ for all $\omega \in B_j$. Writing $\tilde{x}_j = x_{k(j)}$ for each j , we observe that the simple random variable X can be written in an alternative, but equivalent, form $X(\omega) = \sum_{j=1}^m \tilde{x}_j 1_{B_j}(\omega)$. Conversely, it is easy to see that if $X(\omega)$ has a representation of this form with respect to some measurable partition $\{B_j\}$ then it must be the case that $\{B_j\}$ is a refinement of $\{A_k\}$ and, furthermore, $\tilde{x}_j = x_k$ if $B_j \subseteq A_k$.

Now suppose a simple random variable has a representation $X(\omega) = \sum_{j=1}^m \tilde{x}_j 1_{B_j}(\omega)$ with respect to some measurable partition $\{B_j\}$. Does it follow that $E(X) = \sum_{j=1}^m \tilde{x}_j P(B_j)$? (We recall that we had defined the expectation of X with respect to the standard representation (2.1).) To each event A_k , let J_k be the family of indices j for which $B_j \subseteq A_k$. The index family J_k is non-empty as $\{B_j\}$ is itself a partition, hence must cover the set A_k . It follows that $\bigcup_{j \in J_k} B_j = A_k$ and $\tilde{x}_j = x_k$ for every $j \in J_k$. And we find indeed,

$$\begin{aligned} \sum_{j=1}^m \tilde{x}_j P(B_j) &= \sum_{k=1}^n \sum_{j \in J_k} \tilde{x}_j P(B_j) = \sum_{k=1}^n \sum_{j \in J_k} x_k P(B_j) \\ &= \sum_{k=1}^n x_k \sum_{j \in J_k} P(B_j) = \sum_{k=1}^n x_k P\left(\bigcup_{j \in J_k} B_j\right) = \sum_{k=1}^n x_k P(A_k). \end{aligned}$$

Of course, the quantity on the right-hand side is just the expectation of X . Thus, *the expectation of any simple random variable may be defined with respect to any convenient representation, not just the standard representation*. All roads lead to Rome.

The fact that simple expectation is additive and homogeneous is key to all that follows.

THEOREM 1 *If X and Y are simple random variables then $E(X + Y) = E(X) + E(Y)$.*

PROOF: As X and Y are simple, they are of the form $X(\omega) = \sum_{k=1}^n x_k 1_{A_k}(\omega)$ and $Y(\omega) = \sum_{j=1}^m y_j 1_{C_j}(\omega)$. As $\{A_k\}$ and $\{C_j\}$ are both measurable partitions

$1_{A_k}(\omega) = \sum_j 1_{A_k \cap C_j}(\omega)$ and, likewise, $1_{C_j}(\omega) = \sum_k 1_{A_k \cap C_j}(\omega)$. Hence

$$(X + Y)(\omega) = \sum_{k=1}^n \sum_{j=1}^m (x_k + y_j) 1_{A_k \cap C_j}(\omega)$$

and $X + Y$ is also a simple random variable. Consequently, by invariance of expectation to the actual representation,

$$\begin{aligned} E(X + Y) &= \sum_{k=1}^n \sum_{j=1}^m (x_k + y_j) P(A_k \cap C_j) = \sum_{k=1}^n x_k \sum_{j=1}^m P(A_k \cap C_j) \\ &+ \sum_{j=1}^m y_j \sum_{k=1}^n P(A_k \cap C_j) = \sum_{k=1}^n x_k P(A_k) + \sum_{j=1}^m y_j P(C_j) = E(X) + E(Y), \end{aligned}$$

the penultimate step following by total probability. ▶

THEOREM 2 *If X is simple and c any constant then $E(cX) = c E(X)$.*

PROOF: The demonstration is completely banal. If $X = \sum_{k=1}^n x_k 1_{A_k}$ then

$$E(cX) = \sum_{k=1}^n (cx_k) P(A_k) = c \sum_{k=1}^n x_k P(A_k) = c E(X)$$

as cX is clearly also simple. ▶

Additivity and homogeneity together imply that if r is any positive integer, X_1, \dots, X_r any collection of simple random variables, and c_1, \dots, c_r any collection of constants, then $E(\sum_{i=1}^r c_i X_i) = \sum_{i=1}^r c_i E(X_i)$. In other words, *the expectation of simple random variables is a linear operation*.

Additivity of expectation explains why the mean of the binomial is the sum of Bernoulli means as observed at the end of Section VIII.2. So we have two different proofs of the basic result that a binomial with parameters n and p has mean np . Which is better? From the standpoint of generality and applicability, the second approach is to be preferred: while the direct approach is tailored to the particular properties of the binomial, the latter approach uses no combinatorial properties peculiar to the situation but instead relies upon a much more general phenomenon—linearity.

One other possibility is suggested by the binomial. Write $X^{(n)}$ for the number of successes in n tosses of a bent coin with success probability p . Then, as we've seen, $E(X^{(n)}) = np$ for all values of n . For every choice of $m < n$ we then have the trite but suggestive pair of equations $X^{(m)} \leq X^{(n)}$ and $E(X^{(m)}) \leq E(X^{(n)})$. Trivial? Yes, but possibly worthy of remark. And, in point of fact, the utility of this simple observation is far out of proportion to its apparently trivial nature, as we shall see. Deep consequences often can be traced to humble beginnings.

THEOREM 3 Suppose X and Y are simple random variables. If X is dominated by Y , that is to say, $X(\omega) \leq Y(\omega)$ for all ω , then $E(X) \leq E(Y)$.

PROOF: The random variable $Z(\omega) = Y(\omega) - X(\omega)$ is simple and positive. It follows that, in any representation $Z(\omega) = \sum_{i=1}^s \zeta_i 1_{C_i}(\omega)$, all the values ζ_i must be positive (why?). And as all summands in $E(Z) = \sum_i \zeta_i P(C_i)$ are positive, it follows that Z has positive expectation. By additivity it now follows that $0 \leq E(Z) = E(Y - X) = E(Y) - E(X)$ so that $E(Y) \geq E(X)$. ▶

The reader is encouraged to try to prove monotonicity of expectation for simple random variables without appealing to linearity. The proof is instructive and recommended; and once she has it under her belt she will really have understood what a simple random variable entails.

3 Expectations unveiled

In moving from discrete-valued random variables to absolutely continuous variables the transition from an expectation sum with respect to a probability distribution to an expectation integral with respect to a density is so natural as to occasion no comment. But it is not nearly so clear how the definition of expectation can be extended further to cover situations where the distribution is continuous but singular and, more generally, to arbitrary mixture distributions. As the reader has seen in Example XII.4.4, singular continuous distributions can crop up abruptly in apparently innocuous situations.

EXAMPLES: 1) *The singular instance of the gambler's gain.* Suppose Z_1, Z_2, \dots is a sequence of Bernoulli trials corresponding to tosses of a fair coin. Then, as we saw in Example I.7.7, in base 2, the real number $.Z_1 Z_2 \dots = \sum_{k=1}^{\infty} Z_k 2^{-k}$ is uniformly distributed in the unit interval. Now suppose a gambler bets on the results of the even-numbered tosses and, for each k , stands to gain $3 \cdot 2^{-2k}$ if the $2k$ th toss is a success. (The multiplying factor 3 is introduced solely to have things add up prettily.) Writing $X_k = Z_{2k}$ for the result of the $2k$ th trial, the gambler's net winnings are then given by the random variable $X = 3 \sum_{k=1}^{\infty} 4^{-k} X_k$. What is the distribution F of his winnings?

If $X_1 = 1$ then it is clear that $X \geq 3/4$ while if $X_1 = 0$ then $X \leq 3(4^{-2} + 4^{-3} + \dots) = 1/4$. It follows that the interval $(1/4, 3/4)$ is a level set of the distribution function $F(x)$ which takes the value $1/2$ in this range by symmetry (indeed, for any $1/4 < x < 3/4$, $X \leq x$ if, and only if, $X_1 = 0$, an event of probability $1/2$). Proceeding apace, if $X_1 = 0$ and $X_2 = 1$ then $3/16 \leq X \leq 1/4$ while if $X_1 = 0$ and $X_2 = 0$ then $X \leq 3(4^{-3} + 4^{-4} + \dots) = 1/16$, and it follows that $(1/16, 3/16)$ is another level set of F with $F(x) = 1/4$ over this interval as $X < 3/16$ if, and only if, $X_1 = X_2 = 0$; arguing likewise by varying the outcome X_2 when $X_1 = 1$ we discover that $(13/16, 15/16)$ is another level set of F where

the d.f. takes value $3/4$ (as $X < 15/16$ if, and only if, $X_1 = 0$ or $(X_1, X_2) = (1, 0)$). Proceeding recursively in this fashion, at each step a level set constituting an open interval of constancy is carved out of the middle of each of the remaining closed subintervals of the unit interval. Some of the level sets of the d.f. are sketched in Figure 2 and the reader is invited to imagine an endless procession

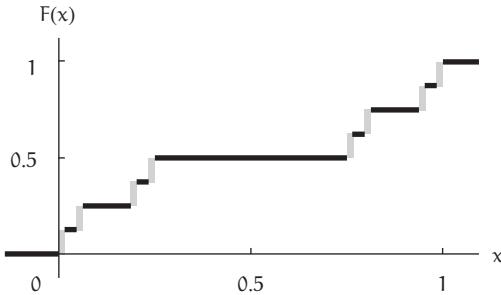


Figure 2: The quaternary Cantor distribution.

of a larger and larger number of tinier and tinier intervals ultimately filling up all the lightly shaded space between the intervals shown. The level sets of F may be enumerated inductively as shown in Table 1 for the first five stages of interval specification and one can readily imagine how to extend the process indefinitely.

The reader will recall that we had encountered a variant of this d.f. before in Example XII.4.4. As we saw there, the resulting Cantor d.f. is continuous but its points of increase are confined to a set of measure zero whence its derivative is zero almost everywhere. Thus, we are confronted with a singular continuous distribution and a density is abeyant.

How much should the gambler have to pay to participate in the game? The game is *fair in the classical sense* if the entrance fee is equal to the expected gain. Here the d.f. $F(x)$ is continuous and, it may be recalled, is built up of a sequence of level sets with all its points of increase confined to an uncountable set (the quaternary Cantor set) of measure zero. In consequence, the d.f. has a derivative which is zero almost everywhere. A naïve approach to expectation by computing a weighted integral with respect to the derivative of $F(x)$ will yield zero. This is clearly not a reasonable value (from the point of view of the gambling house) for the expected gain of the gambler; indeed, a gambling house that charges a zero entrance fee for this game will be very popular for the short period of its existence. The need for a more flexible and powerful definition of expectation is manifest here.

2) *Distributions on the circle.* Singular continuous distributions also arise naturally in settings where variables in two or more dimensions are confined to a

		$\left(\frac{1}{1024}, \frac{3}{1024}\right)$
		$\left(\frac{1}{256}, \frac{3}{256}\right)$
		$\left(\frac{13}{1024}, \frac{15}{1024}\right)$
		$\left(\frac{1}{64}, \frac{3}{64}\right)$
		$\left(\frac{49}{1024}, \frac{51}{1024}\right)$
		$\left(\frac{13}{256}, \frac{15}{256}\right)$
		$\left(\frac{61}{1024}, \frac{63}{1024}\right)$
		$\left(\frac{1}{16}, \frac{3}{16}\right)$
		$\left(\frac{193}{1024}, \frac{195}{1024}\right)$
		$\left(\frac{49}{256}, \frac{51}{256}\right)$
		$\left(\frac{205}{1024}, \frac{207}{1024}\right)$
		$\left(\frac{13}{64}, \frac{15}{64}\right)$
		$\left(\frac{241}{1024}, \frac{243}{1024}\right)$
		$\left(\frac{61}{256}, \frac{63}{256}\right)$
		$\left(\frac{253}{1024}, \frac{255}{1024}\right)$
		$\left(\frac{1}{4}, \frac{3}{4}\right)$
		$\left(\frac{769}{1024}, \frac{771}{1024}\right)$
		$\left(\frac{193}{256}, \frac{195}{256}\right)$
		$\left(\frac{781}{1024}, \frac{783}{1024}\right)$
		$\left(\frac{49}{64}, \frac{51}{64}\right)$
		$\left(\frac{817}{1024}, \frac{819}{1024}\right)$
		$\left(\frac{205}{256}, \frac{207}{256}\right)$
		$\left(\frac{829}{1024}, \frac{831}{1024}\right)$
		$\left(\frac{13}{16}, \frac{15}{16}\right)$
		$\left(\frac{961}{1024}, \frac{963}{1024}\right)$
		$\left(\frac{241}{256}, \frac{243}{256}\right)$
		$\left(\frac{973}{1024}, \frac{975}{1024}\right)$
		$\left(\frac{61}{64}, \frac{63}{64}\right)$
		$\left(\frac{1009}{1024}, \frac{1011}{1024}\right)$
		$\left(\frac{253}{256}, \frac{255}{256}\right)$
		$\left(\frac{1021}{1024}, \frac{1023}{1024}\right)$

Table 1: Level sets of the quaternary Cantor distribution.

lower-dimensional manifold.

Consider an experiment whose outcome corresponds to the selection of a random point on a circle of radius $1/2\pi$. The sample space is formally the coordinate plane (X_1, X_2) with the distribution concentrated on a circle of unit circumference. The uniform distribution on the circle is characterised by the property that for any arc \mathbb{A} of the circle, the probability that the random pair (X_1, X_2) lies in \mathbb{A} is equal to the length of the arc $\ell(\mathbb{A})$. Here is another example where the derivative of the joint distribution function $F(x_1, x_2)$ is zero almost everywhere in the plane. Any continuous function, say, $Y = g(X_1, X_2)$ inherits its distribution from the singular continuous distribution of the pair (X_1, X_2) . Insofar as our current development of the theory is concerned, it is hence problematic to refer the computation of the expectation of Y to the originating singular continuous distribution $F(x_1, x_2)$.

It may be objected that this is a straw man and that what we have here is really just a concealed problem in one dimension. Indeed, in Section IX.2 we had sidestepped the issue by considering problems that depended only on the length of the arc, this contrivance allowing us to effectively reduce the problem to one dimension. But ignoring the situation in two dimensions does not make the difficulty vanish or obviate the need to provide a framework that can handle settings like this. For instance, in Section IX.3 we considered the length of the vector sum of variables each of which is uniformly distributed on a circle and in this case we cannot simply treat the problem as if the circle were the sample space. It is clear that this situation is not limited to these examples but recurs if the variables take values on a torus, or in general, on any lower-dimensional manifold of an n -dimensional space. ▶

A secondary source of inspiration in seeking a general framework for expectation is the annoying discrepancy evidenced in the notation for expectation in the discrete and absolutely continuous cases. A unified theory and notation is a consummation devoutly to be wished.

A third reason, more subtle and compelling, arises when one considers sequences of random variables—situations that arise frequently in theory and practice. The Riemann theory of integration does not really handle passages to the limit well and there is a pressing need for a theory of expectation where limiting situations are handled effortlessly and transparently.

The answer was provided in an elegant general framework for integration established by Henri Lebesgue in an ambitious and far-reaching programme that came to fruition in the year 1902. The key idea seized upon by Lebesgue was that it may be possible to build up a sequence of approximations to a given random variable by a sequence of simple random variables. And if we are very good and lucky (and say our prayers every night) perhaps the limit of the expectations of these simple random variables may be what we are looking for.

THE LEBESGUE PROGRAMME The process begins with a consideration of simple random variables for which we already have a non-controversial, intuitive measure of expectation and then systematically constructs the notion of expectation for more and more complicated random variables using simple random variables as building blocks.

- ① *Simple expectations.* Suppose S is a simple random variable with a representation $S(\omega) = \sum_{k=1}^v s_k 1_{A_k}(\omega)$ in terms of some finite measurable partition $\{A_k, 1 \leq k \leq v\}$. Then *the expectation of S is given by*

$$E(S) = \sum_{k=1}^v s_k P(A_k). \quad (3.1)$$

- ② *Positive expectations.* The key next step in the Lebesgue development defines expectation of positive-valued random variables in terms of the expectations of approximating simple functions. *If $X = X(\omega)$ is a positive-valued random variable, its expectation is defined to be*

$$E(X) = \sup E(S) \quad (3.2)$$

where the supremum is over all simple random variables S dominated by X , that is to say, satisfying $S(\omega) \leq X(\omega)$ for every sample point ω . If the supremum is unbounded we set $E(X) = +\infty$.

- ③ *General expectations.* The hard work is done. A general random variable X may be uniquely represented as a difference of two positive random variables, $X = X^+ - X^-$, where X^+ is the *positive part of X* defined by

$$X^+(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0, \\ 0 & \text{if } X(\omega) < 0, \end{cases}$$

and X^- , likewise, is the *negative part of X* defined by

$$X^-(\omega) = \begin{cases} 0 & \text{if } X(\omega) \geq 0, \\ -X(\omega) & \text{if } X(\omega) < 0. \end{cases}$$

Note that $E(X^+)$ and $E(X^-)$ are both well-defined (if possibly infinite) as X^+ and X^- are both positive random variables. *The expectation of X is now defined by*

$$E(X) = E(X^+) - E(X^-) \quad (3.3)$$

if at least one of $E(X^+)$ and $E(X^-)$ is finite. In this case, X is said to have an expected value (which may be infinite). If both $E(X^+)$ and $E(X^-)$ are finite then X is said to be integrable (or to have finite expectation). If both $E(X^+)$ and $E(X^-)$ are infinite then X does not have expectation.

For a simple random variable $S = \sum_{k=1}^n s_k 1_{A_k}$ the notion of expectation is tied to the intuitive idea of centre of mass, as we've already seen, and does not depend on the particular choice of representation; any old measurable partition where S is constant on each of the sets A_k will do. Expectation here is a finite sum, hence causes no technical difficulties. In consequence, the expectation of simple random variables is well-defined with the connection with physical ideas of centre of mass providing a pleasing intuitive interpretation. Satisfactory on all counts.

The key to a successful generalisation is the canny idea of approximating the expectation of positive random variables by the expectation of simple functions in (3.2). The definition is certainly compact; that it is also remarkably supple and subtle can only be appreciated after a searching examination.

4 Approximation, monotone convergence

A little graphic terminology helps reduce the mathematical verbiage. Suppose U and V are random variables and $U(\omega) \leq V(\omega)$ for all ω . As before, we say then that U is *dominated by* V .

Now suppose X is any positive random variable, $X(\omega) \geq 0$ for all ω . The definition (3.2) says that its expectation $E(X)$ is the supremum of the expectations of simple functions dominated by X . This definition is all very spare and elegant, but two questions come to mind. How does one compute the expectation in practice? And, how does this definition of expectation square with the intuitive notions we've already seen in the case of discrete and absolutely continuous random variables? Consider the following process of approximation of any positive random variable X . For each positive integer m , set $A_{mn} = \{\omega : \frac{n}{2^m} \leq X(\omega) < \frac{n+1}{2^m}\}$ for $0 \leq n \leq m2^m - 1$ and set $A_{m,m2^m} = \{\omega : X(\omega) \geq m\}$. The events A_{mn} partition the range of $X(\omega)$ into finer and finer intervals as m increases, while simultaneously expanding the range covered at each step. These are shown schematically in Figure 3 where the x-axis represents an abstract sample space with the sample points ω as abscissa and the ordinate connotes the values of the function $X(\omega)$. As m increases the reader will notice how the ordinate gets partitioned finer and finer (though in a nod to the exigencies of space, I have not shown how the family of parallel lines keeps encroaching inexorably further and further up the y-axis as m increases). The sets A_{mn} corresponding to given intervals are shown shaded on the x-axis. The reader should observe that for each value of $m \geq 0$ the simple function

$$S_m(\omega) = \sum_{n=0}^{m2^m} \frac{n}{2^m} 1_{A_{mn}}(\omega) \quad (4.1)$$

approximates $X(\omega)$ from below. It is now easy to see that the sequence of simple functions $S_m(\omega)$ satisfies the following properties:

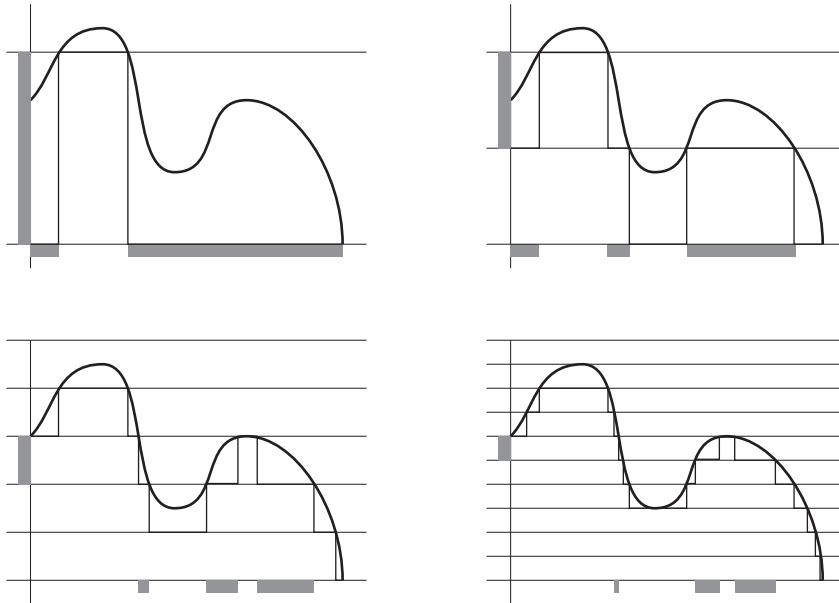


Figure 3: Approximation by simple functions.

- 1) The functions are *positive*, $S_m(\omega) \geq 0$.
- 2) The sequence is *increasing*, $S_m(\omega) \leq S_{m+1}(\omega)$.
- 3) The sequence *converges pointwise* to X , $S_m(\omega) \rightarrow X(\omega)$ for each ω .

It is tempting to conjecture that $E(S_m) \rightarrow E(X)$ as $m \rightarrow \infty$. And in fact it does. The key is a justly celebrated theorem of Beppo Levi.

MONOTONE CONVERGENCE THEOREM Suppose $\{X_n\}$ is an increasing sequence of positive random variables converging pointwise to a random variable X . In picturesque notation, $0 \leq X_n \uparrow X$. Then $\lim E(X_n) = E(X)$.

PROOF: We begin with the observation that if U and V are any two positive random variables with U dominated by V then any simple function dominated by U is certainly also dominated by V . It follows that

$$\begin{aligned} E(U) &= \sup\{E(S) : S \text{ is simple}, S \leq U\} \\ &\leq \sup\{E(S) : S \text{ is simple}, S \leq V\} = E(V). \end{aligned}$$

As, for each n , X_n is dominated by X_{n+1} , it follows that $\{E(X_n), n \geq 1\}$ is an increasing sequence, hence has a limit, possibly infinite. But as each X_n is

dominated by X , we also have $E(X_n) \leq E(X)$ for each n . We conclude therefore that $\lim_{n \rightarrow \infty} E(X_n) \leq E(X)$. To show that $E(X_n) \rightarrow E(X)$ it hence suffices to show that $\lim_{n \rightarrow \infty} E(X_n) \geq E(X)$.

Let us pick any simple random variable S dominated by X (for instance, as given by (4.1)). To give ourselves some room to manoeuvre between S and X , pick any small $\epsilon > 0$ as a slack variable and consider the simple random variable $(1 - \epsilon)S$ instead. Clearly, $(1 - \epsilon)S(\omega) < X(\omega)$ for each ω . The idea behind the proof is to exploit the fact that the functions $X_n(\omega)$ will ultimately squeeze in between $(1 - \epsilon)S(\omega)$ and $X(\omega)$ as $X_n(\omega)$ converges pointwise to $X(\omega)$. A cartoon of what is going on is sketched in Figure 4 with the x -axis connoting the underlying sample space Ω and the dotted line showing the graph of X_n . By choosing better and better simple approximations S and smaller and smaller

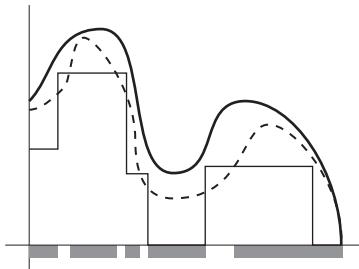


Figure 4: Approximation from below.

slack variables ϵ , we squeeze the limiting expectation of X_n between the expectations of X and a simple random variable whose expectation is arbitrarily close to that of X leaving the limiting expectation of X_n no alternative but to coincide with that of X .

How can we go about making this argument rigorous? We begin with a little notational lubricant. For each n , let A_n be the set of sample points ω at which $X_n(\omega) \geq (1 - \epsilon)S(\omega)$. (These are the points shaded on the x -axis in the figure.) On this set it is clear that X_n dominates $(1 - \epsilon)S$. As the sequence $\{X_n\}$ is increasing, it follows that $(1 - \epsilon)S1_{A_n} \leq X_n \leq X_m$ for each $m \geq n$. Consequently, $(1 - \epsilon)S1_{A_n}$ is a simple function dominated by X_m so that *a fortiori*

$$E(X_m) \geq E((1 - \epsilon)S1_{A_n}) = (1 - \epsilon)E(S1_{A_n})$$

as expectation is homogeneous for simple random variables. As the inequality holds for all $m \geq n$, we obtain

$$\lim_{m \rightarrow \infty} E(X_m) \geq (1 - \epsilon)E(S1_{A_n}). \quad (4.2)$$

Now let us consider what happens if we allow n to tend to infinity. The left-hand side is unaffected. What happens to the right-hand side?

As S is simple, it may be represented with respect to some measurable partition $\{B_j, 1 \leq j \leq \nu\}$ in the form $S(\omega) = \sum_{j=1}^{\nu} s_j 1_{B_j}(\omega)$ whence $(S1_{A_n})(\omega) = \sum_{j=1}^{\nu} s_j 1_{B_j \cap A_n}(\omega)$, and so

$$E(S1_{A_n}) = \sum_{j=1}^{\nu} s_j P(B_j \cap A_n). \quad (4.3)$$

Now $X_n \leq X_{n+1}$ for each n so that the events A_n are also increasing: $A_n \subseteq A_{n+1}$ for every n . Furthermore, as X_n converges pointwise to X , it follows that $\bigcup_n A_n = \Omega$. We may draw two conclusions for each value of j : first, the events $\{B_j \cap A_n, n \geq 1\}$ are increasing, $B_j \cap A_n \subseteq B_j \cap A_{n+1}$, and second, $\bigcup_n (B_j \cap A_n) = B_j$. It follows that $\lim_{n \rightarrow \infty} P(B_j \cap A_n) = P(B_j)$ by the continuity axiom of probability measure. Taking the limit of both sides of (4.3) as $n \rightarrow \infty$,

$$\begin{aligned} \lim_{n \rightarrow \infty} E(S1_{A_n}) &= \lim_{n \rightarrow \infty} \sum_{j=1}^{\nu} s_j P(B_j \cap A_n) \\ &= \sum_{j=1}^{\nu} s_j \lim_{n \rightarrow \infty} P(B_j \cap A_n) = \sum_{j=1}^{\nu} s_j P(B_j) = E(S), \end{aligned}$$

the interchange of sum and limit causing no difficulties as the sum is finite. As (4.2) holds for all values of n , by taking the limit of both sides as $n \rightarrow \infty$, we obtain

$$\lim_{m \rightarrow \infty} E(X_m) \geq (1 - \epsilon) E(S).$$

The above inequality holds for every choice of $\epsilon > 0$ and consequently,

$$\lim_{m \rightarrow \infty} E(X_m) \geq \sup\{(1 - \epsilon) E(S) : \epsilon > 0\} = E(S).$$

As we had placed no restriction on the choice of S (other than it be dominated by X), our refined inequality holds for all choices of simple functions S dominated by X . Accordingly,

$$\lim_{m \rightarrow \infty} E(X_m) \geq \sup\{E(S) : S \text{ is simple}, S \leq X\} = E(X),$$

and thus $\lim E(X_n) = E(X)$, concluding the proof. ►

The theorem can be strengthened to admit increasing sequences that fail to be convergent but only on a negligible set: if $0 \leq X_n(\omega) \uparrow X(\omega)$ a.e. then $E(X_n) \rightarrow E(X)$. The conclusion follows directly from our theorem by virtue of the following observation whose proof the reader may wish to attempt on her own.

EQUIVALENCE LEMMA Suppose X and Y are positive random variables that are equal a.e. Then $E(X) = E(Y)$.

PROOF: Let $Z = \max\{X, Y\}$ and let $S(\omega) = \sum_{j=1}^{\gamma} s_j 1_{A_j}(\omega)$ be any simple function dominated by Z . Suppose \mathfrak{N} is the null set of points on which X and Y differ. Partition each set A_j into the sets $A'_j = A_j \setminus \mathfrak{N}$ and $B'_j = A_j \cap \mathfrak{N}$ and set $s'_j = \inf\{X(\omega) : \omega \in B'_j\}$. Then the simple function $S' = \sum_{j=1}^{\gamma} s'_j 1_{A'_j} + \sum_{j=1}^{\gamma} s'_j 1_{B'_j}$ is dominated by X whence

$$E(X) \geq E(S') = \sum_{j=1}^{\gamma} s_j P(A'_j) + \sum_{j=1}^{\gamma} s'_j P(B'_j) = \sum_{j=1}^{\gamma} s_j P(A_j) + \sum_{j=1}^{\gamma} (s'_j - s_j) P(B'_j)$$

as $P(A'_j) = P(A_j) - P(B'_j)$. But each B'_j is contained in the null set \mathfrak{N} , hence has zero probability. It follows that $E(X) \geq \sum_{j=1}^{\gamma} s_j P(A_j)$. Thus, to each simple function S dominated by Z we can associate a simple function S' dominated by X and such that $E(S') = E(S)$. It follows that $E(X) \geq E(S)$ for every simple function S dominated by Z , whence $E(X) \geq E(Z)$. On the other hand, $E(Z) \geq E(X)$ for trivial reasons. And thus, $E(X) = E(Z)$ identically. An entirely similar argument shows that $E(Y) = E(Z)$ and it follows that X and Y have the same expectation. ▶

The monotone convergence theorem clears the way for effortless passages to the limit—any convergent, increasing sequence of functions will do—and our previously constructed sequence of simple functions (4.1) provides a ready-made starting point. Now

$$\begin{aligned} E(S_m) &= \sum_{n=0}^{m2^m-1} \frac{n}{2^m} P\left\{\frac{n}{2^m} \leq X < \frac{n+1}{2^m}\right\} + m P\{X \geq m\} \\ &= \sum_{n=0}^{m2^m-1} \frac{n}{2^m} [F(2^{-m}(n+1)-) - F(2^{-m}n-)] + m[1 - F(m-)], \end{aligned} \quad (4.4)$$

the arguments of the d.f. F on the right denoting limits from the left in consequence of the intervals of interest here being closed on the left and open on the right. As $0 \leq S_m \uparrow X$, it follows courtesy the monotone convergence theorem that $E(S_m) \rightarrow E(X)$ and the right-hand side provides a sequence of increasingly good approximations to $E(X)$.

The approximating sum (4.4) simplifies to familiar expressions when X is either discrete or absolutely continuous. Let us immediately turn to some examples.

If X is itself simple then it is clear that the relation (3.2) simplifies to (3.1).

EXAMPLE 1) Bernoulli trials. Suppose X is a Bernoulli trial with success probability p . Then X is itself simple, $X(\omega) = 1_A(\omega)$ where A is the set of sample points at which $X(\omega) = 1$. It follows that $E(X) = P(A) = p$, as it must. ▶

The approximating sum (4.4) points the way towards deploying (3.2) when X takes values over a denumerably infinite set of points.

EXAMPLE 2) Return to the Poisson distribution. Suppose X has the Poisson distribution $p(k; \lambda) = e^{-\lambda} \lambda^k / k!$ concentrated on the positive integers. Thus, among

the events A_{mn} in (4.1), only the events $A_{m,k2^m}$ for integer k running from 0 to m have probability > 0 with $P(A_{m,k2^m}) = p(k;\lambda)$ for $0 \leq k \leq m-1$ and $P(A_{m,m2^m}) = \sum_{j=m}^{\infty} p(j;\lambda)$. It follows that

$$E(S_m) = \sum_{k=0}^{m-1} kp(k;\lambda) + m \sum_{j=m}^{\infty} p(j;\lambda) = e^{-\lambda}\lambda \sum_{k=1}^{m-1} \frac{\lambda^{k-1}}{(k-1)!} + e^{-\lambda}m \sum_{j=m}^{\infty} \frac{\lambda^j}{j!}.$$

The first sum on the right is a partial sum of the exponential series, hence converges to e^λ as $m \rightarrow \infty$, while the second sum on the right may be bounded by

$$\sum_{j=m}^{\infty} \frac{\lambda^j}{j!} = \frac{\lambda^m}{m!} \sum_{j=m}^{\infty} \frac{\lambda^{j-m} m!}{j!} < \frac{\lambda^m}{m!} \sum_{j=m}^{\infty} \frac{\lambda^{j-m}}{(j-m)!} = \frac{\lambda^m}{m!} e^\lambda.$$

The simple bound $m! > (m/e)^m$ [see (IV.5.2)] shows that the contribution of the second term to the expectation is hence less than $\frac{\lambda^m}{(m-1)!} < \lambda \left(\frac{\lambda e}{m-1}\right)^{m-1}$ which vanishes asymptotically. It follows that $E(S_m) \rightarrow \sum_{k=0}^{\infty} kp(k;\lambda) = \lambda$. ▶

The general discrete case follows the same pattern. Suppose that X is a positive-valued random variable taking values in a countable set of points, arranged for definiteness as $0 \leq x_1 < x_2 < \dots$, with X taking value x_k with probability p_k . Suppose the series $\sum_{k=1}^{\infty} x_k p_k$ is convergent. Then, for any $\epsilon > 0$, we can find a positive integer $K = K(\epsilon)$ so that $\sum_{k>K} x_k p_k < \epsilon$. Now select m so large that each of the points x_1, \dots, x_K is contained in a unique interval $[\frac{n}{2^m}, \frac{n+1}{2^m})$. The approximating sum (4.4) then reduces to an expression of the form $\sum_{k=1}^K x_k p_k + \zeta$ where $|\zeta| \leq \epsilon + 2^{-m} < 2\epsilon$ eventually. The finite sum differs from $\sum_{k=1}^{\infty} x_k p_k$ by no more than ϵ . Accordingly, $|lim E(S_m) - \sum_{k=1}^{\infty} x_k p_k| < 3\epsilon$, and as this holds for every $\epsilon > 0$ it follows that $lim E(S_m) = \sum_{k=1}^{\infty} x_k p_k$.

If, on the other hand, X takes values in a continuum and has a density then the sum gives way in the usual fashion to a Riemann integral.

EXAMPLE 3) Return to the uniform distribution. Let X be uniformly distributed in the unit interval. As $F(s+t) - F(s) = t$ if $0 < s < s+t < 1$, the approximating sum (4.4) reduces to $E(S_m) = \sum_{n=0}^{2^m-1} \frac{n}{2^m} \cdot \frac{1}{2^m}$. The right-hand side is a Riemann sum and as $m \rightarrow \infty$ it converges to the corresponding Riemann integral, $lim E(S_m) = \int_0^1 x dx = 1/2$. (For the reader who prefers an explicit computation, the arithmetic series on the right yields $E(S_m) = \frac{1}{2} - \frac{1}{2^{m+1}}$ and so $E(S_m) \rightarrow 1/2$ as it must.) ▶

The general absolutely continuous case is not much more complicated. Suppose X has absolutely continuous d.f. $F(x)$ with density $f(x) = dF(x)/dx$ with support over $x \geq 0$. Writing $x = 2^{-m}n$, it becomes clear that the summands on the right of (4.1) accumulate over differential steps,

$$\Delta F = F(2^{-m}(n+1)-) - F(2^{-m}n-) \approx f(x)\Delta x$$

where $\Delta x = 2^{-m}$ connotes an infinitesimal quantity, and it becomes irresistible to replace the Riemann sum by an integral. Accordingly, if $\int_0^\infty xf(x) dx$ is convergent then, for any $\epsilon > 0$, we may replace the sum on the right of (4.4) by $\int_0^\infty xf(x) dx$ incurring an error of no more than ϵ for a sufficiently large choice of m ; the tail term $m P\{X \geq m\}$ likewise is bounded above by $\int_m^\infty xf(x) dx$ which is also eventually forced to be less than ϵ . It follows that $\lim E(S_m)$ differs from $\int_0^\infty xf(x) dx$ by no more than 2ϵ in absolute value and as $\epsilon > 0$ is arbitrary this forces $\lim E(S_m) = \int_0^\infty xf(x) dx$ identically.

The Lebesgue programme hence gives results consistent with the naïve theory of expectation. The theory is however already more general and flexible and allows us, among other things, to handle singular continuous distributions within the same framework.

EXAMPLE 4) Gambler's gain, continued. The gambler's gain X in Example 3.1 has a singular continuous distribution $F(x)$ whose points of increase have measure zero. How to proceed to compute the expected value? When all else fails, we turn to first principles and consider approximations of the gain X by simple functions à la (4.1). As the gambler's gain is bounded between 0 and 1 we may consider approximations by simple functions also bounded in this range. Now an examination of Table 1 shows that the level sets subdivide available subintervals into fourths at each stage, suggesting a sequence of approximating simple functions of the form

$$S_m(\omega) = \sum_{n=0}^{4^m} \frac{n}{4^m} 1_{A_{mn}}(\omega) \quad (m \geq 1)$$

where the slivers A_{mn} are defined by $A_{mn} = \left\{ \frac{n}{4^m} \leq X < \frac{n+1}{4^m} \right\}$ for $0 \leq n \leq 4^m - 1$ and $A_{m,4^m} = \{X \geq 1\}$. (This is only a slight modification of the sequence specified in (4.1) obtained by discarding the odd-numbered terms in the sequence and renumbering.) It is clear that $\{S_m\}$ is an increasing sequence of simple functions that converges pointwise to the gambler's gain X . Now

$$E(S_m) = \sum_{n=0}^{4^m} \frac{n}{4^m} P\left\{ \frac{n}{4^m} \leq X < \frac{n+1}{4^m} \right\} = \sum_{n=0}^{4^m} \frac{n}{4^m} \Delta F\left(\frac{n}{4^m}\right).$$

The sum is simple to evaluate as $\Delta F(n/4^m)$ is identically zero if the n th interval $[n/4^m, (n+1)/4^m]$ is contained within a level set of the d.f.; else $\Delta F(n/4^m) = 1/2^m$ identically. Systematically identifying the level sets (the reader may find it helpful to refer to the table and the figure) we obtain

$$E(S_1) = \left(\frac{0}{4} + \frac{3}{4}\right) \frac{1}{2} = \frac{3}{8},$$

$$E(S_2) = \left(\frac{0}{16} + \frac{3}{16} + \frac{12}{16} + \frac{15}{16}\right) \frac{1}{4} = \frac{15}{32},$$

$$E(S_3) = \left(\frac{0}{64} + \frac{3}{64} + \frac{12}{64} + \frac{15}{64} + \frac{48}{64} + \frac{51}{64} + \frac{60}{64} + \frac{63}{64}\right) \frac{1}{8} = \frac{63}{128},$$

$$\begin{aligned} E(S_4) = & \left(\frac{0}{256} + \frac{3}{256} + \frac{12}{256} + \frac{15}{256} + \frac{48}{256} + \frac{51}{256} + \frac{60}{256} + \frac{63}{256} + \frac{192}{256} \right. \\ & \left. + \frac{195}{256} + \frac{204}{256} + \frac{207}{256} + \frac{240}{256} + \frac{243}{256} + \frac{252}{256} + \frac{255}{256} \right) \frac{1}{16} = \frac{255}{512}. \end{aligned}$$

Further terms in the sequence may be obtained along these lines with a little good will and patience but the pattern is already visible: $E(S_m) = \frac{4^m - 1}{2 \cdot 4^m}$ for every $m \geq 1$. (The reader may wish to try her hand at an inductive proof if she does not find the exhibited pattern compelling.) The monotone convergence theorem now readily allows us to conclude that $E(X) = \lim_{m \rightarrow \infty} E(S_m) = \frac{1}{2}$. Thus, the game has a fair entrance fee of $1/2$ as is reasonable and intuitive. ►

The hard work is done. To finish up, suppose X is a general random variable and let X^+ and X^- be its positive and negative parts, respectively. Then X^+ and X^- are individually positive random variables whence the expectations $E(X^+)$ and $E(X^-)$ are well-defined via (3.2), if possibly infinite. The expectation of X defined by (3.3), $E(X) = E(X^+) - E(X^-)$, is now well-defined provided that at least one of the two expectations on the right is finite.

We mention for completeness that the theory may now be extended to *complex-valued* random variables with little effort. A complex-valued function of the sample space $Z(\omega) = X(\omega) + iY(\omega)$ is a complex random variable if $X(\omega)$ and $Y(\omega)$ are ordinary, common or garden-variety real random variables. Here, of course, $i = \sqrt{-1}$ is the imaginary root of unity. *The expectation of the complex random variable $Z = X + iY$ is now defined by $E(Z) = E(X) + iE(Y)$ provided both expectations on the right exist.*

To close the loop, is this formulation of expectation completely consistent with the naïve theory for random variables which may, in general, have positive and negative parts? Suppose X is discrete and takes values in the set $\{x_k\}$ with corresponding probabilities $\{p_k\}$. Let $\{x_k^+\}$ denote the subset of positive values in the set $\{x_k\}$, with $\{p_k^+\}$ the corresponding probabilities and, likewise, let $\{x_k^-\}$ denote the subset of negative values in the set $\{x_k\}$ with $\{p_k^-\}$ the corresponding probabilities. Then X^+ is positive and takes values in $\{x_k^+\} \cup \{0\}$ with the same corresponding probabilities $\{p_k^+\}$ with the exception of the point 0 for which the probability sees an increment equal to $P\{X < 0\}$. The expectation of X^+ as we have seen will then coincide with the expression thrown up by the naïve theory, $E(X^+) = \sum_k x_k^+ p_k^+$. (The reader should verify that the point $\{0\}$ contributes nothing to the expectation.) Likewise, X^- is positive and discrete and takes values in $\{-x_k^-\} \cup \{0\}$ with corresponding probabilities $\{p_k^-\}$ again excepting only the point $\{0\}$ which has probability $P\{X \geq 0\}$. The expectation of X^- is hence given by the sum $E(X^-) = \sum_k (-x_k^-) p_k^-$, with $\{0\}$ again contributing nothing to the expectation. If X has expectation it then follows that

$$E(X) = E(X^+) - E(X^-) = \sum_k x_k^+ p_k^+ - \sum_k (-x_k^-) p_k^- = \sum_k x_k p_k$$

exactly as defined in the naïve theory.

Similarly, if X is absolutely continuous with density f , an almost identical argument to the one above shows that, if X has expectation, then

$$E(X) = E(X^+) - E(X^-) = \int_0^\infty xf(x) dx - \int_{-\infty}^0 (-x)f(x) dx = \int_{-\infty}^\infty xf(x) dx$$

again in accordance with the naïve theory. Very satisfactory.

The stage is set. We now need to confirm that the familiar properties of expectation are preserved in this flexible framework.

5 Arabesques of additivity

At the elementary level, the additivity of expectation is natural, even trite. If, for instance, X and Y are arithmetic and possessed of the joint distribution $\{p_{ij}\}$ then

$$E(X + Y) = \sum_{i,j} (i + j)p_{ij} = \sum_{i,j} ip_{ij} + \sum_{i,j} jp_{ij} = E(X) + E(Y)$$

provided, of course, that the expectations exist. If X and Y are absolutely continuous with joint density $f(x, y)$ then, with the usual proviso on existence,

$$E(X + Y) = \iint (x + y)f(x, y) dxdy = \iint xf(x, y) dydx + \iint yf(x, y) dxdy$$

by linearity of integration, and again $E(X + Y) = E(X) + E(Y)$.

Most of what follows needs only an appreciation of this basic property and the reader who is willing to take the demonstration that additivity holds in general on trust can proceed blithely to the next section to be regaled by applications.

To verify that the claimed additivity holds in general one has to return to the flexible and general definition of Section 3. Already the waters appear to be muddied. If, for instance, X and Y are positive, then on the face of it, the definition $E(X) = \sup E(S)$ over all simple random variables S dominated by the positive X only guarantees *super-additivity*, to wit, $E(X + Y) \geq E(X) + E(Y)$ for all positive X and Y . (The family of simple functions dominated by $X + Y$ contains the family of simple functions of the form $S + T$ where S and T are simple and dominated by X and Y , respectively.) The fact that equality holds is nothing short of remarkable and argues a very high degree of prescience in the originator of the definition. (A little bit of luck couldn't have hurt either.)

THEOREM 1 *Suppose X and Y are integrable. Then $E(X + Y) = E(X) + E(Y)$.*

PROOF: We've already seen that expectation is additive for simple random variables. Now suppose X and Y are positive. Let $\{X_n\}$ and $\{Y_n\}$ be increasing

sequences of simple random variables converging pointwise to X and Y , respectively. Then $\{X_n + Y_n\}$ is an increasing sequence of simple random variables converging pointwise to $X + Y$. Three applications of the monotone convergence theorem show hence that $E(X_n) \rightarrow E(X)$, $E(Y_n) \rightarrow E(Y)$, and $E(X_n + Y_n) \rightarrow E(X + Y)$. On the other hand, expectation is additive for simple random variables, so that $E(X_n + Y_n) = E(X_n) + E(Y_n)$ for each n . It follows that the limits must coincide, $E(X + Y) = E(X) + E(Y)$, and expectation is additive for positive random variables.

Now suppose X and Y are arbitrary integrable random variables and consider $Z = X + Y$. Showing additivity of expectation would be a simple matter if one could claim that $Z^+ = (X + Y)^+ = X^+ + Y^+$ but this is in general untrue. A small reworking, however, smooths out all difficulties. We have the decompositions $X + Y = (X + Y)^+ - (X + Y)^-, X = X^+ - X^-$, and $Y = Y^+ - Y^-$. It follows that $(X + Y)^+ + X^- + Y^- = (X + Y)^- + X^+ + Y^+$. Both sides are sums of positive random variables and by the additivity result just shown,

$$E((X + Y)^+) + E(X^-) + E(Y^-) = E((X + Y)^-) + E(X^+) + E(Y^+),$$

or, after rearranging terms,

$$E((X + Y)^+) - E((X + Y)^-) = (E(X^+) - E(X^-)) + (E(Y^+) - E(Y^-)).$$

But the left-hand side is identically $E(X + Y)$ while the right-hand side is identified with $E(X) + E(Y)$. ▶

As an immediate corollary, the reader should observe that X is *integrable* (*i.e.*, has finite expectation) if, and only if, the positive random variable $|X|$ has finite expectation. Indeed, $|X| = X^+ + X^-$ so that, by additivity, $E(|X|) = E(X^+) + E(X^-)$. Thus, for $|X|$ to have finite expectation it is necessary and sufficient that $E(X^+)$ and $E(X^-)$ are both finite. And, of course, the latter condition is necessary and sufficient for $X = X^+ - X^-$ to have finite expectation. As an immediate consequence, we have the useful *modulus inequality* for expectation,

$$|E(X)| = |E(X^+) - E(X^-)| \leq E(X^+) + E(X^-) = E(|X|),$$

assuming, of course, that X has expectation.

What is remarkable about additivity of expectation is that no matter how complex the dependency structure between random variables, as evidenced in their joint distribution, the expectation of the sum is completely determined by the marginal distributions. In particular, the random variables need not be independent.

To establish that the expectation operation is linear it now suffices to show that it is homogeneous. As usual, this is the easy part of linearity.

THEOREM 2 Suppose X is integrable and c any real constant. Then $E(cX) = cE(X)$.

PROOF: The dance is always the same. We first verify the result for simple random variables, then for positive random variables via the monotone convergence theorem, and finally for general random variables.

We've already seen that expectation is homogeneous for simple random variables. Now suppose c is a positive constant and X a positive random variable. If $\{X_n\}$ is an increasing sequence of positive, simple random variables converging pointwise to X then $\{cX_n\}$ is an increasing sequence of positive, simple random variables converging pointwise to cX . Then $E(X_n) \rightarrow E(X)$ and $E(cX_n) \rightarrow E(cX)$. But expectation of simple random variables is homogeneous so that $E(cX_n) = cE(X_n)$. We conclude, as we must, that $E(cX) = cE(X)$.

Now suppose X is any integrable random variable and c is positive. Then $cX = (cX)^+ - (cX)^- = cX^+ - cX^-$ and

$$E(cX) = E(cX^+) - E(cX^-) = cE(X^+) - cE(X^-) = cE(X)$$

by the just-demonstrated homogeneity of expectation for positive random variables. If c is negative apply the same argument to $cX = (cX)^+ - (cX)^- = (-c)X^- - (-c)X^+$ to obtain

$$\begin{aligned} E(cX) &= E((-c)X^-) - E((-c)X^+) = (-c)E(X^-) - (-c)E(X^+) \\ &= cE(X^+) - cE(X^-) = cE(X). \end{aligned}$$

It follows that expectation is homogeneous in general. ▶

Additivity and homogeneity together establish the basic property that expectation is a *linear operation*: if X_1, \dots, X_n are integrable random variables and c_1, \dots, c_n are arbitrary real numbers then $E(\sum_{k=1}^n c_k X_k) = \sum_{k=1}^n c_k E(X_k)$. The proof follows by a simple exercise in induction.

Linearity of expectation says that the order of expectation and finite summation may be interchanged provided the individual random variables are integrable. The interchange in order of expectation and summation carries over almost painlessly to infinite sums courtesy the monotone convergence theorem though we do have to pay attention to convergence.

THEOREM 3 Let $\{X_k, k \geq 1\}$ be a sequence of integrable random variables defined on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\{c_k, k \geq 1\}$ be a sequence of real constants. If the series $\sum_{k=1}^{\infty} |c_k| E(|X_k|)$ converges then $\sum_{k=1}^{\infty} c_k X_k(\omega)$ converges to a limiting random variable $X(\omega)$ for ω a.e. and $E(X) = \sum_{k=1}^{\infty} c_k E(X_k)$.

PROOF: We may as well assume that the c_k are all identically equal to one (else consider the variables $X'_k = c_k X_k$ instead). We first prove the result for positive random variables X_k . Let $Y_n = \sum_{k=1}^n X_k$ for each $n \geq 1$. Clearly, $\{Y_n\}$ is an increasing sequence of positive random variables so that, by the monotone convergence theorem, $E(\lim_{n \rightarrow \infty} Y_n) = \lim_{n \rightarrow \infty} E(Y_n)$. The left-hand side is

equal to $E(\sum_{k=1}^{\infty} X_k)$ while, by linearity of expectation over finite sums, the right-hand side is equal to $\lim_{n \rightarrow \infty} \sum_{k=1}^n E(X_k) = \sum_{k=1}^{\infty} E(X_k)$ which is finite by hypothesis. It follows that $E(\sum_{k=1}^{\infty} X_k) = \sum_{k=1}^{\infty} E(X_k)$ is finite. We conclude, *a fortiori*, that the series $\sum_{k=1}^{\infty} X_k(\omega)$ converges for almost all ω . Indeed, let \mathfrak{N} denote the set of sample points ω on which the series $\sum_{k=1}^{\infty} X_k(\omega)$ is not convergent. Then, for any $M > 0$, the simple function $M1_{\mathfrak{N}}(\omega)$ is dominated by $\sum_{k=1}^{\infty} X_k(\omega)$. It follows by definition of expectation of positive random variables as the supremum of expectations of all dominated simple random variables that $M P(\mathfrak{N}) = M E(1_{\mathfrak{N}}) \leq E(\sum_{k=1}^{\infty} X_k)$ for every $M > 0$. But the right-hand side is finite so that the result can hold for all $M > 0$ if, and only if, $P(\mathfrak{N}) = 0$, or in other words, the series $\sum_{k=1}^{\infty} X_k(\omega)$ converges for almost all ω to some random variable, say, $X(\omega)$. And thus, $X(\omega)$ is integrable with expectation $E(X) = \sum_{k=1}^{\infty} E(X_k)$.

We've established the theorem when $\{X_k, k \geq 1\}$ is a sequence of positive random variables. To show that the result holds for a general sequence of random variables $\{X_k\}$, we write $X_k = X_k^+ - X_k^-$ and apply the just-proved result to the positive sequences $\{X_k^+\}$ and $\{X_k^-\}$ separately. ▶

Positive variables provide a pleasing application of additivity—with a useful identity as a by-product when the variables are arithmetic.

THEOREM 4 Suppose X is an integrable, positive random variable. Then

$$\sum_{n=1}^{\infty} P\{X \geq n\} \leq E(X) < 1 + \sum_{n=1}^{\infty} P\{X \geq n\}.$$

If, additionally, X is arithmetic then $E(X) = \sum_{n=1}^{\infty} P\{X \geq n\}$.

PROOF: As n ranges through the positive integers the sequence of events $A_n = \{n \leq X < n+1\}$ partition the entire sample space. Accordingly, $X = \sum_{n=0}^{\infty} X 1_{A_n}$ and additivity of expectation yields $E(X) = \sum_{n=0}^{\infty} E(X 1_{A_n})$. It follows that

$$\sum_{n=0}^{\infty} n P(A_n) \leq E(X) < \sum_{n=0}^{\infty} (n+1) P(A_n) = 1 + \sum_{n=0}^{\infty} n P(A_n). \quad (5.1)$$

For each $N \geq 1$, we have $\sum_{n=0}^N n P(A_n) = \sum_{n=0}^N n [P\{X \geq n\} - P\{X \geq n+1\}]$. By a rearrangement of the terms of the partial sums of the series we may express the right-hand side in the form

$$\sum_{n=1}^N [n - (n-1)] P\{X \geq n\} - N P\{X \geq N+1\} = \sum_{n=1}^N P\{X \geq n\} - N P\{X \geq N+1\}.$$

Proceeding to the limit as $N \rightarrow \infty$, we hence obtain

$$\sum_{n=0}^{\infty} n P(A_n) = \sum_{n=1}^{\infty} P\{X \geq n\},$$

the integrability of X guaranteeing via (5.1) that the series is convergent. (The reader who is curious about why the integrability of X implies that $\sum_{n=1}^{\infty} n P\{X \geq n\} \rightarrow 0$ will find the answer in Example 8.3.) If X takes positive integer values only then $X = n$ identically in A_n and the rest of the proof follows as before. ▶

EXAMPLES: 1) *An urn problem.* An urn contains r red and b blue balls. Balls are removed (randomly) from the urn, one at a time, without replacement. What is the expected number of balls that have been removed when the first blue ball is taken out? As the reader is well aware, sampling without replacement creates niggling dependencies; the previous result, however, allows us to preserve our sang froid. The setting is that of Pólya's urn scheme of Section II.5. Let X be the number of balls drawn until the first blue ball is drawn. Then the event $X > n$ occurs if, and only if, the first n balls drawn are red. Accordingly,

$$\begin{aligned} P\{X > n\} &= \frac{r}{b+r} \cdot \frac{r-1}{b+r-1} \cdots \frac{r-n+1}{b+r-n+1} \\ &= \frac{r^n}{(b+r)^n} = \binom{b+r-n}{r-n} / \binom{b+r}{r} \quad (0 \leq n \leq r). \end{aligned}$$

It follows that

$$E(X) = \sum_{n \geq 0} P\{X > n\} = \frac{1}{\binom{b+r}{r}} \sum_{n=0}^r \binom{b+r-n}{r-n} = \frac{1}{\binom{b+r}{r}} \sum_{k=0}^r \binom{b+k}{k}.$$

It is irresistible to apply Pascal's triangle to the terms of the sum on the right and we obtain

$$\sum_{k=0}^r \binom{b+k}{k} = \sum_{k=0}^r \left[\binom{b+k+1}{k} - \binom{b+k}{k-1} \right] = \binom{b+r+1}{r}$$

as the sum telescopes. It follows that

$$E(X) = \binom{b+r+1}{r} / \binom{b+r}{r} = \frac{b+r+1}{b+1} = 1 + \frac{r}{b+1},$$

an intuitively satisfying result.

2) *The mean of the exponential.* If X is continuous, positive, and integrable then Theorem 4 has the continuous analogue $E(X) = \int_0^\infty P\{X > t\} dt$ which I will leave to the reader as an exercise. If X is exponentially distributed with parameter α then $P\{X > t\} = e^{-\alpha t}$ for every $t \geq 0$, whence $E(X) = \int_0^\infty e^{-\alpha t} dt = 1/\alpha$. This example is provided purely for purposes of illustration as the reader can justly remark that the direct path to the computation of the mean of the exponential is hardly any more complicated. ▶

Our proof of the monotone convergence theorem explicitly used the fact that monotonicity is inherent in the way expectation was defined for positive random variables. Indeed, if X and Y are positive, $X \leq Y$, then the family of simple functions dominated by Y certainly includes the family of simple functions dominated by X and consequently $E(X) \leq E(Y)$. The proof for general random variables follows from additivity along the usual lines.

THEOREM 5 *Suppose X and Y are integrable and $X \leq Y$. Then $E(X) \leq E(Y)$.*

PROOF: Suppose X and Y are integrable random variables with Y dominating X . As $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ it follows that $X^+ + Y^- \leq X^- + Y^+$. Both sides of the inequality are positive random variables whence

$$E(X^+) + E(Y^-) = E(X^+ + Y^-) \leq E(X^- + Y^+) = E(X^-) + E(Y^+).$$

Recombining terms, we obtain $E(X^+) - E(X^-) \leq E(Y^+) - E(Y^-)$, or, what expresses the same thing, $E(X) \leq E(Y)$. ▶

Thus, *expectation preserves inequalities*. In particular, we obtain the trite but useful *modulus inequality* $E(|X|) \leq E(|Y|)$. And, as another easy consequence of the proof, if X is dominated by some integrable random variable Y then X has finite expectation.

Monotonicity is trite and innocuous looking but has deep consequences. We will explore some of these in Chapter XVII.

6 Applications of additivity

Additivity weaves a subtle dance through the theory of expectation and in retrospect is the single most remarkable and important feature of the theory. An application in a familiar setting provides an unsubtle but convenient starting point for exploration.

EXAMPLE 1) Waiting times, revisited. What is the expected number of failures before the r th success in a succession of coin tosses? Our analysis of the negative binomial suggests that a linearity property is at work. To see the origins of this phenomenon more clearly, break up the problem into a succession of smaller problems as follows: let Y_1 denote the number of failures before the first success; Y_2 the number of failures between the first success and the second; Y_3 the number of failures between the second success and the third; and so on. In general, Y_k is the number of failures between the $(k-1)$ th success and the k th success. Then the number of failures before r successes are garnered is clearly given by $X = Y_1 + Y_2 + \dots + Y_r$. Each Y_k has a geometric distribution¹ with

¹Indeed, though we will not need the fact in this example, a little introspection shows that Y_1, \dots, Y_r are *independent and identically distributed* random variables. The phenomenon illustrated here is a *recurrent event*.

parameter p and, as we've seen, $E(Y_k) = q/p$. Linearity of expectation does the rest and $E(X) = \sum_{k=1}^r E(Y_k) = rq/p$, as obtained earlier. ►

Additivity of expectation is particularly potent in situations where the summands are dependent. We begin with problems in a discrete setting.

EXAMPLES: 2) *The hat check problem, revisited.* If n individuals are matched with their hats at random, what is the expected number of individuals who get their own hats? Denote by A_k the event that the k th individual is matched with his own hat. Write $X = X_n$ for the number of individuals who get their hats. Then $X_n = \sum_{k=1}^n 1_{A_k}$. The probability that the k th individual is matched with his own hat is exactly $(n-1)!/n! = 1/n$ by symmetry. It follows that the indicator 1_{A_k} has mean $1/n$. Consequently, $E(X_n) = \sum_{k=1}^n E(1_{A_k}) = n \cdot \frac{1}{n} = 1$ for every n . The reader will observe that in this case the events A_1, \dots, A_n (*a fortiori* the corresponding indicators) are *dependent*. No matter. Linearity of expectation is unaffected by the dependency structure of the constituent random variables in the sum.

The reader should compare this with the direct approach: writing P_m for the probability that exactly m individuals get their hats we obtain $E(X_n) = \sum_{m=0}^n mP_m = \sum_{m=1}^n mP_m$. We can readily write an explicit expression for P_m . For $k = 1, \dots, n$, we observe that

$$\sigma_k = \sum_{1 \leq i_1 < \dots < i_k \leq n} P(A_{i_1} \cap \dots \cap A_{i_k}) = \binom{n}{k} \frac{1}{n^k} = \frac{1}{k!}.$$

Inclusion-exclusion now readily yields

$$P_m = \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} \sigma_{m+k} = \sum_{k=0}^{n-m} (-1)^k \binom{m+k}{m} \frac{1}{(m+k)!} = \sum_{k=0}^{n-m} \frac{(-1)^k}{m!k!}.$$

It follows that

$$E(X_n) = \sum_{m=1}^n m \left(\sum_{k=0}^{n-m} \frac{(-1)^k}{m!k!} \right) = \sum_{j=0}^{n-1} \sum_{k=0}^{n-j-1} \frac{(-1)^k}{j!k!}.$$

It may not be immediately obvious that the double sum above is identically equal to 1 for every value of n . The reader who tries her hand at it will have discovered a pretty combinatorial problem and, incidentally, rediscovered what linearity told us in one easy swoop.

3) *The birthday problem, revisited.* The birthday paradox provides another discrete setting where dependencies are rife. What is the expected number of days in a year that are birthdays to *exactly* k individuals in a given population of n individuals? Suppose a year is comprised of exactly $N = 365$ days and that

each individual has a birthday distributed uniformly across the year. Let A_j denote the event that the j th day of the year is host to exactly k birthdays in the population, and let 1_{A_j} denote its indicator. Then the number of days that are birthdays of exactly k individuals is given by $X = 1_{A_1} + \dots + 1_{A_N}$ so that

$$E(X) = E(1_{A_1}) + \dots + E(1_{A_N}) = Nb_n(k; 1/N) = \binom{n}{k} N^{-n+1} (N-1)^{n-k}.$$

Here's another situation where the individual events A_j are dependent. The reader will more fully appreciate the power of linearity if she attempts a direct combinatorial attack on the problem instead.

4) *An urn problem, revisited.* Additivity suggests another approach to the urn problem of Example 5.1. If we consider the balls selected sequentially, each blue interrupts a sequence (possibly null) of reds. Thus, we obtain $b+1$ red sequences of lengths N_1, N_2, \dots, N_{b+1} (some of which may be zero). The sum of these lengths must equal r , $N_1 + N_2 + \dots + N_{b+1} = r$. On the other hand, by symmetry, each of these sequence lengths has the same (marginal) distribution whence, taking expectation, $r = E(N_1) + E(N_2) + \dots + E(N_{b+1}) = (b+1) E(N_1)$. (The variables N_j are clearly dependent. No matter! Expectation is always additive.) It follows that the expected number of contiguous reds drawn before the first blue is $r/(b+1)$; including the first blue drawn finishes up. ►

We turn to the familiar uniform and normal densities to furnish continuous examples; the Cantor distribution provides an outlet from the humdrum.

EXAMPLES: 5) *Random points.* Suppose that X_1, \dots, X_n are random points in the unit interval in the sense of Section IX.1 and let $Z = \min\{X_1, \dots, X_n\}$. What can be said about $E(Z)$?

A direct attack on the distribution of Z is not difficult but the reader may find the following alternative tack even more appealing. With probability one the X_k divide the unit interval into $n+1$ subintervals of lengths L_1, \dots, L_{n+1} . Here we may identify the length L_1 of the first subinterval with Z .

If we now imagine the origin as an $n+1$ th point X_0 and "roll up" the unit interval so that the points $X_0 = 0$ and 1 coincide then the points X_0, X_1, \dots, X_n are random points on a unit circumference circle and it becomes apparent that the $n+1$ intervals L_1, \dots, L_{n+1} are the arc lengths between these points. The symmetry of the problem forces the L_k to share a common marginal distribution though they are of course highly dependent. As $L_1 + \dots + L_{n+1} = 1$, additivity of expectation yields $E(L_1) + \dots + E(L_{n+1}) = (n+1) E(L_1) = 1$ whence $E(L_1) = E(Z) = 1/(n+1)$. A little thought should show the reader that this is just the continuous analogue of the previous example.

6) *Correlated normals.* Suppose X_1 and X_2 are jointly normal random variables with density $f(x_1, x_2) = \frac{1}{\sigma_1 \sigma_2} \phi\left(\frac{x_1 - \mu_1}{\sigma_1}, \frac{x_2 - \mu_2}{\sigma_2}; \rho\right)$ for arbitrary μ_1 and μ_2 , positive σ_1 and σ_2 , and $|\rho| < 1$. Here $\phi(\xi_1, \xi_2; \rho)$ is the bivariate normal density given in (VII.5.3). Suppose $Y = X_1 + X_2$. What is the mean of Y ?

The marginal densities of X_1 and X_2 are both normal and given by $f_1(x_1) = \frac{1}{\sigma_1} \phi(\frac{x_1 - \mu_1}{\sigma_1})$ and $f_2(x_2) = \frac{1}{\sigma_2} \phi(\frac{x_2 - \mu_2}{\sigma_2})$, respectively, as seen in Example VII.5.5. It follows that $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$, and, consequently, $E(Y) = \mu_1 + \mu_2$. A direct attack on the problem would require one to first determine the distribution of Y (it's normal with mean $\mu = \mu_1 + \mu_2$ and variance $\sigma^2 = \sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2$ but this needs to be shown), followed by the evaluation of the mean of Y from its density. Additivity of expectation, however, required only the computation of the marginal (normal) distribution of the constituent random variables X_1 and X_2 .

7) *Return to the Cantor distribution.* Suppose $X = 3 \sum_{k=1}^{\infty} X_k 4^{-k}$ where $\{X_k\}$ is a sequence of Bernoulli trials each taking values 0 or 1, each with probability 1/2. The distribution of X is the singular continuous distribution of Example 3.1 and we can proceed as in Example 4.4 to compute its expectation. Here's an alternative path which the reader may find slightly more appealing.

As $E(X_k) = 1/2$ for each k , the series

$$3 \sum_{k=1}^{\infty} E(X_k) 4^{-k} = \frac{3}{2} \sum_{k=1}^{\infty} 4^{-k} = \frac{3}{2} \cdot \frac{4^{-1}}{1 - 4^{-1}} = \frac{1}{2}$$

is convergent and hence, by linearity of expectation, $E(X) = 1/2$. ▶

The power of additivity is particularly apparent in the following abstract setting.

EXAMPLE 8) Another waiting time average. Suppose X_1, \dots, X_n are independent and identically distributed random variables taking only positive values and suppose the expectations $E(X_k)$ and $E(X_k^{-1})$ both exist (and are finite). Let $S_n = X_1 + \dots + X_n$. What can be said about $E(X_k S_n^{-1})$? (The finicky reader will worry that it is first necessary to guarantee the existence of $E(S_n^{-1})$. But this is immediate. Why?)

A direct attack would require the determination of the joint distribution of X_k and S_n (note that they are dependent random variables) followed by the evaluation of an expectation integral in two dimensions. A much more direct path is available, however, starting from the elementary observation

$$1 = \frac{S_n}{S_n} = \frac{X_1 + \dots + X_n}{S_n} = \frac{X_1}{S_n} + \dots + \frac{X_n}{S_n}.$$

Taking expectations of both sides shows hence via linearity that

$$1 = E\left(\frac{X_1}{S_n}\right) + \dots + E\left(\frac{X_n}{S_n}\right).$$

But the random variables X_1, \dots, X_n are independent with a common distribution whence it follows by symmetry that $E(X_k S_n^{-1})$ is a constant (for given n) that does not depend on k . It follows that $E(X_k S_n^{-1}) = \frac{1}{n}$ for each $k = 1, \dots, n$. ▶

These examples may have gone some way towards convincing the reader of the potency of additivity. A slogan would not be amiss.

SLOGAN *The expectation of a sum is the sum of expectations.*

Our next application requires a little preparation, the result of import enough to merit a section all to itself. The reader will find deeper applications of additivity scattered through the rest of the book.

7 The expected complexity of Quicksort

We say that a set \mathbb{A} is *totally ordered* if there exists a binary relation (traditionally denoted \leq) on \mathbb{A} such that for any two elements a and b in \mathbb{A} we have $a \leq b$ or $b \leq a$, both relations holding simultaneously if, and only if, a and b coincide, and, furthermore, if any three elements a , b , and c in \mathbb{A} satisfy $a \leq b$ and $b \leq c$ then $a \leq c$. Thus, any two elements of a totally ordered set are comparable (hence *total order*) and the elements of the set may be linearly ordered by the transitivity property. It is irresistible to think of the set as being ordered by “size” and we do so even though the elements may not be numerical.

Examples of basic totally ordered sets include the letters of the alphabet ordered by their dictionary order $a \leq b \leq c \leq \dots$, the natural numbers, the integers, the rational numbers, and the real numbers. Totally ordered sets engender new sets imbued with a total order in very natural ways. If \mathbb{B} is a totally ordered set and $f: \mathbb{A} \rightarrow \mathbb{B}$ is an injection then \mathbb{A} is totally ordered by setting $a \leq b$ when $f(a) \leq f(b)$. If we specialise this observation to the case when \mathbb{A} is a subset of \mathbb{B} then by identifying the injection f with the identity map $f: a \mapsto a$ we have the natural consequence that every subset of a totally ordered set inherits a total order from its parent. Thus, the natural numbers, the integers, and the rationals all inherit their total order from the real numbers. More complex totally ordered sets can be systematically built up. If \mathbb{W} is the Cartesian product of a countable family $\{\mathbb{A}_j, j \geq 1\}$ of totally ordered sets then the lexicographic ordering of the elements of \mathbb{W} induces a total order: if $a = (a_1, a_2, \dots)$ and $b = (b_1, b_2, \dots)$ are distinct elements of \mathbb{W} then there is a smallest index, say, k , for which $a_k \neq b_k$; the lexicographic ordering then sets $a \leq b$ if $a_k \leq b_k$ and $b \leq a$ if $b_k \leq a_k$. Thus, if we begin with a finite alphabet \mathbb{A} and expand it by the addition of a space symbol (which is smaller than all the letters) then any set of words ordered lexicographically is a totally ordered set. The Index of this book provides a familiar example in this class.

Suppose a_1, \dots, a_n is a finite collection of elements from a totally ordered set. These could represent, for instance, a list of names or topics, or a spreadsheet of numbers. In many such situations it is desirable to sort these elements alphabetically, lexicographically, or in order of size. The aforementioned Index provides a milieu where an (alphabetic) sorting of entries is desirable.

The reader will be able to come up with other examples from day-to-day life: team, class, and group lists, payrolls, rankings, and order statistics all come readily to mind. Before the advent of digital computers, manipulating data in applications like these had to be done laboriously by hand and sorting beyond a few tens of entries was extremely time consuming. Today, of course, sorting needs beyond the most rudimentary are handled effortlessly and transparently deus ex machina; the reader is probably familiar, for instance, with the sorting feature that is common in electronic spreadsheets which, with the click of a button, sorts large lists painlessly, transparently, and apparently instantaneously. But the faster computation of a silicon brain only means that it can handle larger databases before coming unstuck. To maximise the gains from a faster processor we will need to compute *efficiently*. This leads us to ask the following question: given a finite collection of elements a_1, \dots, a_n from a totally ordered set, how can we efficiently order them from smallest to largest?

Taking a cue from how we may approach the problem visually, one could begin by scanning the sequence to identify the smallest element, say, a_{j_1} . Remove it from the collection and identify the smallest element, say, a_{j_2} , of the expurgated sequence. Remove it and proceed in this fashion until all the elements in the original sequence are exhausted. The procedure has then unearthed the ordering $a_{j_1} \leq a_{j_2} \leq \dots \leq a_{j_n}$. Let us call this the *naïve procedure*. What is its complexity? Identifying the smallest element in a collection of m elements requires $m - 1$ comparisons. As we begin with a collection of n elements and eliminate one at a time, the complexity of the naïve procedure is given by the arithmetic series² $(n-1) + (n-2) + \dots + 2 + 1 = (n-1)n/2$. The quadratic rate of growth blows up quickly—for $n = 10^6$ the naïve procedure requires about 5×10^{11} comparisons—and we cast about for a better alternative.

The Quicksort algorithm takes a rather different tack to the problem. It begins with an initial group of elements $\mathcal{P} = \{a_1, \dots, a_n\}$ to be sorted. The first step consists of the selection of a random element X from \mathcal{P} and collecting the remaining elements into two disjoint groups \mathcal{P}' and \mathcal{P}'' where \mathcal{P}' consists of elements a with $a \leq X$ and \mathcal{P}'' consists of elements a with $X \leq a$. The selection of X then induces a quasi-ordering $\mathcal{P}' X \mathcal{P}''$ with X properly positioned with respect to the eventual ordering of the sets on either side. The problem is now reduced to a consideration of sets \mathcal{P}' and \mathcal{P}'' , each of cardinality strictly less than \mathcal{P} (either may be empty), and we may proceed recursively. The considerations are the same on either side of X and we consider the left side for definiteness. If $\text{card } \mathcal{P}' = \emptyset$ then we have nothing left to do with this group. If $\text{card } \mathcal{P}' \geq 1$ the

²The arithmetic series formula has been known for a long time. Faced with an unruly class in 1784 the teacher, J. G. Büttner, assigned them the task of adding up all the integers from 1 to 100. While his classmates struggled with it, the young Carl Friedrich Gauss (he was only seven at that time!) produced the answer $5050 = 50 \times 101$ without skipping a beat. Gauss had observed that the integers from 1 to 100 could be grouped into 50 pairs $(1, 100), (2, 99), \dots, (50, 51)$ each pair adding up to 101. It takes genius to transport the pedestrian associative property to magic.

selection of a random element X' from \mathcal{P}' partitions the remaining elements in \mathcal{P}' into two groups, those smaller than X' and those larger than X' , and we have a quasi-ordering to the left of X with the location of X' determined. We proceed systematically in this fashion until all groups are empty, the elements that have been selected falling into the desired increasing order. The description takes more real estate than the execution and an example makes all clear.

EXAMPLE: The sequence of steps shown in Table 2 illustrates how the Quicksort algorithm may proceed for a given sequence of numbers to be arranged in increasing order of size. In each step the selected number from a group is shown

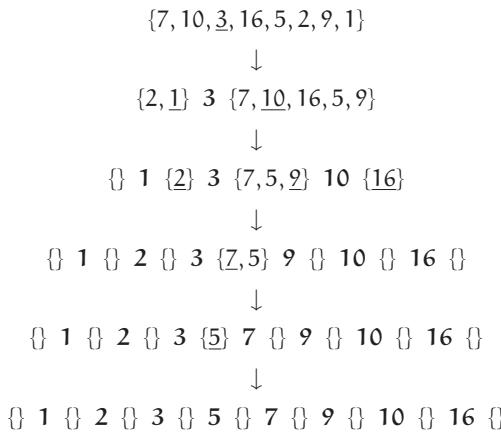


Table 2: The Quicksort algorithm.

underlined, and appears in bold font on the following line indicating its position in the order is now fixed; spaces between groups have been added for legibility. An examination of the steps shows that seven comparisons are needed in step one, five comparisons for the two groups in step two, two comparisons for the single group of size more than one in step three, and one comparison in step four for a grand total of 15 comparisons. The naïve algorithm, on the other hand, requires 28 comparisons and the improvement is already clear. ►

The number of comparisons needed by Quicksort depends on the particular choices of the elements selected in each group. In the worst case its performance is the same as the naïve algorithm—this happens if the Quicksort algorithm selects an extremal member of each group at each stage. The randomisation over the choices prevents systemic inefficiencies from building up because of the occasional poor selection.

An analysis of the expected complexity of Quicksort may be begun by the observation that the algorithm requires at most one comparison between

any two elements. As the algorithm only relies upon the relative order of the elements and not what they are, we may as well identify the elements with the natural numbers $1, \dots, n$. Write X_{jk} for the indicator of the event that a comparison was made between elements j and k . Then the total number of comparisons made by Quicksort is given by $C = \sum_{k=2}^n \sum_{j=1}^{k-1} X_{jk}$, the sum counting the contribution from any given pair of elements exactly once. The determination of the distribution of C is complicated by the dependencies that build across the indicator variables X_{jk} but additivity of expectation cuts right to the chase and we only need to determine $E(X_{jk})$ for $1 \leq j < k \leq n$.

The key to the understanding of whether j is ever compared with k is the block of $k-j+1$ elements $j, j+1, \dots, j+(k-j)$. These elements will all reside as a block in a group until the first time that one of them is selected. If one of the end elements j or $k = j + (j - k)$ is selected then a comparison between j and k is effected; if, however, one of the $k-j-1$ elements between j and k are selected then j and k will be placed in different groups separated by the selected element and can thereafter never be directly compared. Thus, the probability that j is compared with k is equal to the probability that j or k is selected *given that one of $j, j+1, \dots, j+(k-j)$ is selected*. By the uniform choice mechanism, it follows that $E(X_{jk}) = P\{j \text{ is compared with } k\} = 2/(k-j+1)$ and thus

$$E(C) = \sum_{k=2}^n \sum_{j=1}^{k-1} \frac{2}{k-j+1} = \sum_{k=2}^n \sum_{i=2}^k \frac{2}{i}$$

by a change in the inner summation variable to $i = k-j+1$. To get a feeling for the asymptotic rate of growth, we may approximate the inner sum by integrals to obtain

$$\int_2^k \frac{dx}{x} < \sum_{i=2}^k \frac{1}{i} < \int_1^k \frac{dx}{x}.$$

The bookend integrals are easy to evaluate and, as $\sum_{k=2}^n \log k = \log n!$, by consolidating terms we have $2 \log n! - 2(n-1) \log 2 < E(C) < 2 \log n!$. The bounds do not simplify further but Stirling's formula $n! \sim \sqrt{2\pi n^{n+1/2}} e^{-n}$ (for two derivations, each in a rather different spirit, see the theorem of Section XIV.7 and Theorem XVI.6.1) provides a clean asymptotic picture.

THEOREM *As $n \rightarrow \infty$, the expected number of comparisons made by Quicksort is asymptotic to $2n \log n$, that is to say, $E(C)/(2n \log n) \rightarrow 1$.*

Thus, when $n = 10^6$, the expected number of comparisons required by Quicksort is about 3×10^7 which improves over the naïve procedure by four orders of magnitude. In the best case the Quicksort algorithm halves the size of each active group at each step—I will leave it to the reader to verify that

this results in a number of comparisons asymptotic to $n \log_2 n$. Thus, the randomisation over group elements in Quicksort introduces an overhead of a multiplicative factor of only $2/\log 2 \approx 2.885$ over the best result possible. Quicksort was developed by C. A. R. Hoare in 1960.³

Sorting and searching are central features of algorithms and it is not surprising that they should have received a lot of attention. Donald Knuth has devoted an entire volume of his definitive series on the *Art of Computer Programming* to this subject and the reader desirous of learning more can do no better than to consult it.⁴

8 Expectation in the limit, dominated convergence

At some level one could take the position that the primary rôle of the Lebesgue theory in probability is to facilitate smooth passages to the limit. And, indeed, with a little of that trusting faith of which the poet speaks so highly, most of this book can be read without reference to the theory of measure and integration. The advanced theory, however, does require an understanding of the foundational rôle played by measure and the dominated convergence principle is one of the pillars on which the theory rests.

The monotone convergence theorem applies to increasing sequences of random variables. In applications, however, sequences are usually not so obligingly increasing and it would be convenient to have a slightly more flexible tool at hand.

Suppose $\{X_n, n \geq 1\}$ is a positive sequence of integrable random variables. While the sequence may not converge pointwise to some limiting random variable, the quantity $\liminf_n X_n$ certainly exists and represents a, possibly defective, random variable. What can be said about its expectation? The surprisingly useful Fatou's lemma points the way.

FATOU'S LEMMA Suppose $\{X_n, n \geq 1\}$ is a positive sequence of random variables on some probability space. Then $E(\liminf_{n \rightarrow \infty} X_n) \leq \liminf_{n \rightarrow \infty} E(X_n)$.

PROOF: For each n , set $Y_n = \inf\{X_m : m \geq n\}$. Then Y_n is positive and increasing, $Y_n \leq Y_{n+1}$ for each n , so that, by monotone convergence,

$$\lim_{n \rightarrow \infty} E(Y_n) = E\left(\lim_{n \rightarrow \infty} Y_n\right) = E\left(\lim_{n \rightarrow \infty} \inf\{X_m : m \geq n\}\right) = E\left(\liminf_{n \rightarrow \infty} X_n\right).$$

On the other hand, $Y_n \leq X_m$ for all $m \geq n$ whence, by monotonicity of expectation, $E(Y_n) \leq E(X_m)$ for all $m \geq n$. Consequently, $E(Y_n) \leq \inf\{E(X_m) : m \geq n\}$, whence $\lim_{n \rightarrow \infty} E(Y_n) \leq \liminf_{n \rightarrow \infty} E(X_n)$ by passing to the limit. ►

³C. A. R. Hoare, "Quicksort", *Computer Journal*, vol. 5, no. 1, pp. 10–16, 1962.

⁴D. Knuth, *The Art of Computer Programming. Volume 3: Sorting and Searching*. New York: Addison-Wesley, 1973.

Fatou's lemma admittedly looks awkward at first blush and only experience will convince the reader of its utility. Lebesgue's dominated convergence principle, for instance, is an immediate consequence of the lemma and will serve to provide an example of how the lemma eases passages to limits.

DOMINATED CONVERGENCE THEOREM *Suppose $X_n \rightarrow X$ pointwise. If there exists an integrable random variable Y such that $|X_n| \leq Y$ for each n then X is integrable, $E(|X|) \leq E(Y)$, and $\lim_{n \rightarrow \infty} E(X_n) = E(X)$.*

PROOF: Again by monotonicity of expectation, $E(|X_n|) \leq E(Y)$ for each n , whence $\liminf_n E(|X_n|) \leq E(Y)$. Fatou's lemma applied to the positive sequence $|X_n|$ shows that $E(\liminf_n |X_n|) \leq \liminf_n E(|X_n|) \leq E(Y)$. As X_n converges pointwise to X , it follows that $|X_n|$ converges pointwise to $|X|$ and the term on the left is equal to $E(|X|)$. Thus X is integrable and $E(|X|) \leq E(Y)$.

We may convert the problem into a consideration of sequences of positive random variables by defining $U_n = Y - X_n$ and $V_n = Y + X_n$. Fatou's lemma applied to the positive sequence U_n gives

$$\begin{aligned} E(\liminf_n U_n) &\leq \liminf_n E(U_n) = \liminf_n E(Y - X_n) \\ &= \liminf_n (E(Y) - E(X_n)) = E(Y) - \limsup_n E(X_n). \end{aligned}$$

(The reader should remark on the casual appeal to additivity.) But $\liminf_n X_n = \limsup_n X_n = \lim_n X_n = X$, and so

$$E(\liminf_n U_n) = E(Y - \limsup_n X_n) = E(Y - X) = E(Y) - E(X).$$

It follows that $E(X) \geq \limsup_n E(X_n)$. Similarly, Fatou's lemma applied to the positive sequence V_n gives

$$\begin{aligned} E(\liminf_n V_n) &\leq \liminf_n E(V_n) = \liminf_n E(Y + X_n) \\ &= \liminf_n (E(Y) + E(X_n)) = E(Y) + \liminf_n E(X_n) \end{aligned}$$

by additivity. The left-hand side, on the other hand, is given by

$$E(\liminf_n V_n) = E(Y + \liminf_n X_n) = E(Y + X) = E(Y) + E(X),$$

so that we obtain $E(X) \leq \liminf_n E(X_n)$. Combining the two bounds on $E(X)$ we obtain $E(X) \leq \liminf_n E(X_n) \leq \limsup_n E(X_n) \leq E(X)$ from which it follows that $\lim_n E(X_n)$ exists and equals $E(X)$. ▶

As in the case of the monotone convergence theorem, the conditions may be relaxed: *if $X_n(\omega) \rightarrow X(\omega)$ a.e. and $|X_n(\omega)| \leq Y(\omega)$ a.e. where Y is integrable, then $E(X_n) \rightarrow E(X)$* . The demonstration requires little more than to step through the proof again keeping the equivalence lemma in mind. I will leave it to the reader to verify the details.

The name of the theorem arises from the *domination condition* $|X_n| \leq Y$ for each n for some integrable Y . This is perhaps the only place in the theory where a naïve passage to a limit without a glance at the attendant domination condition can lead to trouble, as the following salutary examples show.

EXAMPLES: 1) *A standard construction.* A fertile source of counterexamples is to consider random variables which take large values in small intervals. Suppose the sample space is the unit interval $(0, 1)$ equipped with Lebesgue measure. For each positive integer n , let $X_n(\omega) = n$ if $0 < \omega < 1/n$ and 0 otherwise. It is easy to see that $\lim_{n \rightarrow \infty} X_n(\omega) = 0$ pointwise for every ω . But, on the other hand, $E(X_n) = n \int_0^{1/n} d\omega = 1$ for every n from which it follows that $\lim_{n \rightarrow \infty} E(X_n) = 1$. If we had set $X_n(\omega) = n^2 1_{(0,1/n)}(\omega)$ then X_n again tends to zero pointwise but $E(X_n) = n \rightarrow \infty$. The key point is that there exists no integrable Y that uniformly dominates the X_n .

2) *A geometric construction.* Again consider as sample space the unit interval $(0, 1)$ equipped with the usual notion of length and let $U(\omega) = \omega$ be a uniformly distributed random variable on this space. For each positive integer n , define $X_n = (n+1)(n+2)U^n(1-U)$. For each sample point ω , $X_n(\omega) \rightarrow 0$ as $n \rightarrow \infty$ so that X_n converges pointwise to 0 identically. It is tempting hence to conclude that $\lim_n E(X_n) = 0$. However,

$$E(X_n) = (n+1)(n+2) E(U^n(1-U)) = (n+1)(n+2) [E(U^n) - E(U^{n+1})].$$

As U is uniformly distributed over the unit interval, $E(U^k) = \int_0^1 u^k du = \frac{1}{k+1}$ for any positive k . Thus, $E(X_n) = 1$ for every n , whence $\lim_n E(X_n) = 1$, and the naïve passage to the limit has led to an erroneous conclusion. The sticking point again is that the variables X_n cannot be uniformly dominated by some integrable Y . The reader may find some fun and profit in figuring out why. ►

In typical situations when the dominated convergence theorem comes into play the dominating random variable is usually easy to identify.

EXAMPLE 3) Continuity. Suppose X is an integrable random variable. If $\{A_n\}$ is a sequence of events such that $\lim_{n \rightarrow \infty} P(A_n) = 0$ then $Y_n = X 1_{A_n}$ is a sequence of random variables dominated by $|X|$ and converging pointwise to zero a.e. It follows that $\lim_n E(Y_n) = E(\lim_n X 1_{A_n}) = 0$. In particular, with $A_n = \{|X| \geq n\}$, we have $\lim_n \int_{\{|X| \geq n\}} X dP = 0$, and *a fortiori*, $\lim_n n P\{|X| \geq n\} = 0$. ►

The reader may have wondered why the distribution tail $m P\{X \geq m\}$ on the right of (4.4) does not contribute to the expectation of the integrable X in the limit as $m \rightarrow \infty$. Our example provides the answer. Our proof of Theorem 5.4 provides another application where this tail estimate is needed.

The dominated convergence theorem comes into play when we can exhibit a convergent sequence of random variables. Say we are merely given an increasing sequence of random variables without a limit clearly in evidence. Under what conditions can we assert that the sequence actually converges?

LEVI'S THEOREM Suppose $\{X_n\}$ is an increasing sequence of random variables and there exists M such that $E(X_n) \leq M$ for each n . Then there exists an integrable random variable X such that $X_n \rightarrow X$ a.e. and, moreover, $E(X_n) \rightarrow E(X)$.

PROOF: We may suppose $\{X_n, n \geq 1\}$ to be a sequence of positive random variables by replacing X_n by $X_n - X_1$ and we so suppose. We edge up to the set A of points ω on which $X_n(\omega) \rightarrow \infty$ by first considering the events $A_n^{(r)} = \{\omega : X_n(\omega) \geq r\}$ for any integer $r \geq 1$. Then, for any n , $P(A_n^{(r)}) = E(1_{A_n^{(r)}}) \stackrel{(i)}{\leq} E\left(\frac{1}{r}X_n1_{A_n^{(r)}}\right) \stackrel{(ii)}{\leq} \frac{1}{r}E(X_n) \leq \frac{M}{r}$, the step labelled (i) following because $X_n(\omega)/r \geq 1$ whenever $1_{A_n^{(r)}}(\omega)$ is not identically zero, and (ii) following by homogeneity of expectation and the obvious inequality $0 \leq X_n1_{A_n^{(r)}} \leq X_n$. In step (i) the reader may recognise with pleasure the idea of Chebyshev that we encountered in the proof of the weak law of large numbers in Section V.6; we shall see the idea deployed repeatedly. As the sets $\{A_n^{(r)}, n \geq 1\}$ constitute an increasing sequence, it follows that the set $A^{(r)} = \bigcup_n A_n^{(r)}$ on which one or more of the events $X_n \geq r$ occur has probability also bounded above by M/r . The sets $\{A^{(r)}, r \geq 1\}$ constitute a decreasing sequence with limit $A = \bigcap_r A^{(r)} = \{\omega : X_n(\omega) \rightarrow \infty\}$. Consequently, $P(A) \leq M/r$ for each r , which is to say that $P(A) = 0$ identically.

For any selection $\omega \notin A$, the sequence $\{X_n(\omega), n \geq 1\}$ is increasing and bounded, hence converges to a finite limit, say, $X(\omega)$. On A we may set $X(\omega)$ to any value, say, zero. We've hence established that there exists a random variable X such that $X_n(\omega) \uparrow X(\omega)$ excepting only on the null set A , that is to say, $X_n \uparrow X$ a.e. To show that $E(X_n) \rightarrow E(X)$ it will suffice courtesy the dominated convergence theorem to exhibit an integrable random variable Y that dominates X . The simplest construction suffices here.

For $r \geq 1$, let $B^{(r)}$ denote the set of sample points ω for which $r-1 \leq X(\omega) < r$ and set $Y(\omega) = \sum_{r=1}^{\infty} r1_{B^{(r)}}(\omega)$. Then $X < Y \leq X+1$ and *a fortiori* each X_n is dominated by Y a.e. To conclude that $E(X_n) \rightarrow E(X)$ by the dominated convergence theorem we only need show now that $E(Y) = \lim_s \sum_{r=1}^s rP(B^{(r)})$ is finite. Write $C^{(s)} = \bigcup_{r=1}^s B^{(r)}$ for the event $0 \leq X < s$. Then $E(Y1_{C^{(s)}}) \leq E((X+1)1_{C^{(s)}}) \leq E(X1_{C^{(s)}}) + 1$, by repeated application of additivity and monotonicity. Now $X_n \uparrow X$ a.e. and, in particular, $X_n(\omega) \uparrow X(\omega) < s$ a.e. on $C^{(s)}$. By the dominated convergence theorem, it follows that $E(X1_{C^{(s)}}) = \lim_n E(X_n1_{C^{(s)}})$. But, as $X_n1_{C^{(s)}} \leq X_n$, by monotonicity of expectation, $E(X_n1_{C^{(s)}}) \leq E(X_n) \leq M$ for each n and so also for the limit as $n \rightarrow \infty$. Putting the pieces together we obtain $\sum_{r=1}^s rP(B^{(r)}) \leq M + 1$ and, proceeding to the limit as $s \rightarrow \infty$, we see that Y is integrable. ▶

If $\{Z_k, k \geq 1\}$ is a sequence of integrable positive variables and the series $\sum_k E(Z_k)$ is convergent then by a consideration of the sequence of partial sums $X_n = \sum_{k=1}^n Z_k$ we obtain the specialisation of Levi's theorem used in Section V.7.

9 Problems

1. Balls and urns. An urn contains r red balls and b blue balls. Balls are removed sequentially from the urn (without replacement). What is the expected number of balls left in the urn at the *first* instant at which all the remaining balls are of the same colour?

2. Suppose r balls are distributed randomly in n urns. Show that the expected number of empty urns is $n(1 - \frac{1}{n})^r$. [Hint: Use the recurrence (XII.11.1).]

3. *Expectation.* By refining the proof of Theorem 5.4, show that if X is an integrable, positive random variable then $E(X) = \int_0^\infty P\{X > x\} dx$.

4. Let F be a continuous d.f. and, for each positive integer n , let $X^{(n)}$ be a random variable with d.f. $F(x)^n$. For each positive integer k form the random variable $Y_{k,n} = F(X^{(n)})^k$. Evaluate $E(Y_{k,n})$. [Hint: Consider Problem XII.18.]

5. Suppose X_1, X_2, \dots are independent, positive random variables with a common distribution and suppose both $E(X_1) = a$ and $E(X_1^{-1}) = b$ exist. For each n , let $S_n = X_1 + \dots + X_n$. Evaluate $E(S_m/S_n)$ for $m \leq n$.

6. *Return to the Cantor d.f.* In Example 4.4 we had asserted $E(S_m) = (4^m - 1)2^{m-1}/(4^m 2^m)$. Prove it. Now evaluate $\int_0^1 x^2 dF(x)$ and $\int_0^1 e^{itx} dF(x)$.

7. Show that if X is positive and $E(X1_A) = 0$ then $X = 0$ a.e. on A . By working separately with the positive and negative parts of X , conclude that if $E(X1_A) = 0$ for all $A \in \mathcal{F}$ then $X(\omega) = 0$ a.e.

8. If X is an integrable random variable then, for every $\epsilon > 0$, there exists $\delta > 0$ such that $E(X1_A) < \epsilon$ for every event A of probability no larger than δ .

9. Suppose c is a strictly positive constant. Then X is integrable if, and only if, $\sum_{n=1}^\infty P\{|X| \geq cn\} < \infty$. If the series on the left converges for any value of c then it converges for all values of c .

10. *Approximation by simple functions.* Given an integrable random variable X and $\epsilon > 0$ show that there exists a simple random variable Z_ϵ such that $E(|X - Z_\epsilon|) < \epsilon$. Hence show that there exists a sequence of simple random variables $\{X_n\}$ such that $\lim_n E(|X - X_n|) = 0$, and, furthermore, we may choose $\{X_n\}$ so that $|X_n| \leq |X|$ for all n .

11. *Construction of new probability measures.* Suppose X is a positive, integrable random variable on a probability space (Ω, \mathcal{F}, P) with strictly positive expectation $E(X) > 0$. Define the set function Q on the σ -algebra \mathcal{F} via $Q(A) = E(X1_A)/E(X)$ for each A in \mathcal{F} . Show that Q is a probability measure on Ω .

12. Suppose $X_1 \geq X_2 \geq \dots \geq X_n \geq \dots$ is a decreasing sequence of integrable random variables and suppose that there exists M such that $E(X_n) \geq M$ for each n . There exists an integrable random variable X with $X_n \downarrow X$ a.e. and $E(X_n) \rightarrow E(X)$.

Problems 13–23 deal with a generalisation of the notion of conditional expectation introduced in Section VII.6. Suppose X is a random variable measurable with respect to a σ -algebra \mathcal{F} and \mathcal{G} is any σ -algebra contained in \mathcal{F} , that is to say, \mathcal{G} is a sub- σ -algebra of \mathcal{F} .

DEFINITION A conditional expectation of X given \mathcal{G} , denoted $E(X | \mathcal{G}) = E(X | \mathcal{G})(\omega)$, is a random variable measurable with respect to \mathcal{G} which satisfies $E(E(X | \mathcal{G})1_A) = E(X1_A)$ for every $A \in \mathcal{G}$. If Y is any other random variable on the same space then, by identifying $\mathcal{G} = \sigma(Y)$ with the σ -algebra generated by Y , we write $E(X | Y)$ for $E(X | \sigma(Y))$ and call it (slightly inaccurately but more graphically) a conditional expectation of X given Y .

The random variable X is in general not measurable with respect to \mathcal{G} ; else we could simply take $E(X | \mathcal{G})$ to be X itself. The existence of conditional expectation is assured (by the

Radon–Nikodým theorem of analysis) but it is unfortunately only identified a.e., hence “a” instead of “the” in the definition. I will not detour to prove the Radon–Nikodým theorem here but, naturally, I hope the reader will be willing to assume existence for the purposes of the problems to follow. If she prefers not to take it on faith she will find an elegant, essentially *geometric*, construction of conditional expectation which by-passes the need for the Radon–Nikodým theorem in Problems XIV.46–51. In the following problems variables are measurable with respect to \mathcal{F} and \mathcal{G} is a sub- σ -algebra of \mathcal{F} . The reader should assume integrability where needed.

13. A characteristic property of conditional expectation. Suppose $X = \sum_j x_j 1_{B_j}$ and $Y = \sum_k y_k 1_{A_k}$ are simple random variables. Write $p(j | k) = P\{X = x_j | Y = y_k\}$. Show that $E(X | Y)(\omega) = \sum_k \sum_j x_j p(j | k) 1_{A_k}(\omega)$ is a conditional expectation in the sense of Section VII.6 and hence exhibits the following characteristic property: if A is any element of the (finite) σ -algebra $\sigma(Y)$ induced by Y , then $E(X 1_A) = E(E(X | Y) 1_A)$. This discovery motivates the preceding abstract formulation of conditional expectation.

14. Show that if $\mathcal{G} = \mathcal{F}$ then we may take $E(X | \mathcal{G}) = X$ and if $\mathcal{G} = \{\emptyset, \Omega\}$ then we may take $E(X | \mathcal{G}) = E(X)$.

15. Show that conditional expectation is linear and monotone: for any real constants a and b , we have $E(aX + bY | \mathcal{G}) = aE(X | \mathcal{G}) + bE(Y | \mathcal{G})$ a.e.; if $X \leq Y$, then $E(X | \mathcal{G}) \leq E(Y | \mathcal{G})$ a.e.

16. Continuation, corollary. If $X \geq 0$, then $E(X | \mathcal{G}) \geq 0$ a.e.

17. Tower property. If $\mathcal{G}_1 \subseteq \mathcal{G}_2$ (\mathcal{G}_1 is coarser than \mathcal{G}_2) then $E(E(X | \mathcal{G}_2) | \mathcal{G}_1) = E(X | \mathcal{G}_1)$ a.e. If $\mathcal{G}_1 \supseteq \mathcal{G}_2$ (\mathcal{G}_1 is finer than \mathcal{G}_2) then $E(E(X | \mathcal{G}_2) | \mathcal{G}_1) = E(X | \mathcal{G}_2)$ a.e.

18. Total probability. Show that $E(E(X | \mathcal{G})) = E(X)$.

19. Independence. If X [more precisely, $\sigma(X)$] is independent of \mathcal{G} then $E(X | \mathcal{G}) = E(X)$ a.e.

20. Factorisation property. If W is measurable with respect to \mathcal{G} and W and WX are both integrable, then $E(WX | \mathcal{G}) = W E(X | \mathcal{G})$ a.e.

21. Monotone convergence. If $0 \leq X_n \uparrow X$, then $E(X_n | \mathcal{G}) \uparrow E(X | \mathcal{G})$ a.e.

22. Fatou. If $X_n \geq 0$, then $E(\liminf X_n | \mathcal{G}) \leq \liminf E(X_n | \mathcal{G})$ a.e.

23. Dominated convergence. Suppose $X_n \rightarrow X$ a.e. If there exists an integrable random variable Y such that $|X_n(\omega)| \leq Y(\omega)$ for all n , then $E(X_n | \mathcal{G}) \rightarrow E(X | \mathcal{G})$ a.e.

While integrals are defined only for measurable functions, the concept of an outer integral is occasionally useful in non-measurable cases.

24. Outer integrals. If f is positive, the outer integral of f , denoted $\int^* f dP$, is defined as the infimum of the expectations $\int g dP$ as $g \geq f$ varies over all extended real-valued, \mathcal{F} -measurable functions dominating f . Interpret $\infty - \infty = \infty$ and for arbitrary $f = f^+ - f^-$ define $\int^* f dP = \int^* f^+ dP - \int^* f^- dP$. Let $P^*(A) = \inf\{P(B) : A \subseteq B, B \in \mathcal{F}\}$ denote outer measure. Show that $\int^* 1_A dP = P^*(A)$ for all $A \subseteq \Omega$.

25. Continuation, subadditivity. Show that outer integrals are subadditive: if $f \geq 0$ and $g \geq 0$, then $\int^*(f + g) dP \leq \int^* f dP + \int^* g dP$. Inequality can be strict.

26. Continuation, monotone convergence. If $0 \leq f_n \uparrow f$ then $\int^* f_n dP \uparrow \int^* f dP$.

XIV

Variations on a Theme of Integration

The additivity and monotonicity properties of expectation are reminiscent of integration and, indeed, the two notions are intimately connected. The benefits of a viewpoint of expectation as an abstract integral go beyond ease of notation. We are grown accustomed to writing and manipulating integrals and identifying expectations with integrals creates a link with the familiar and opens new doors of thought.

C 1–5, 10
A 6, 8, 9, 11
7

1 UTILE ERIT SCRIBIT \int PRO OMNIA

Thus famously wrote Gottfried Wilhelm Leibniz in his study in Paris on October 29, 1675. Translated, the phrase becomes “It is useful to write \int for sum”. The selection was inspired—few choices of notation have had as profound an impact on mathematics as did Leibniz’s epiphany.

Notation is a powerful vehicle for focusing thought: the proper choices clarify and occasionally suggest lines of attack; poor choices muddy the waters and obscure the essence of the problem. Leibniz had originally used the word *omnia* for sum, the letters *a* and *l* standing for fixed increments, what, today, we would call the differential quantities dx and dy . Thus, with the short form *omn.* for *omnia*, the expression *omn. l = y* would stand for the indefinite integral $\int dy = y$, while the identity *omn. xl = x omn. l – omn. omn. la* may be more recognisable in the integration by parts formula $\int x dy = xy - \int y dx$. Leibniz’s contraction was much much more than a mere rewriting of a cumbersome formula. His streamlined notation for a sum not only allows us to focus on the essentials by clearing the notational brush, it even suggests a seductive new possibility—that of considering the quantities dx and dy as infinitesimal—leading to a powerful extension of the notion of summation and the birth of a new mathematics.

Leibniz’s words ring true in our setting as well. The expectation of a random variable $(\Omega, \mathcal{F}, \mathbf{P}) \xrightarrow{X} (\mathbb{R}, \mathcal{B}, \mathcal{F})$ is a functional determined implicitly both by the probability measure and the measurable map on the space. One

may view the expectation through the prism of the originating space or the induced space. Both viewpoints lead to a useful consolidation.

THE VIEW FROM $(\Omega, \mathcal{F}, \mathbf{P})$

The rôle played by the underlying space in the expectation becomes clear when one considers the approximation of a positive random variable by simple functions. Suppose X is positive. For each m , partition the range of values of X into $0 = x_{m,0} < x_{m,1} < \dots < x_{m,m2^m-1} < x_{m,m2^m} = m$. This partition of the range of X into tiny intervals induces a corresponding partition of the sample space into “infinitesimal slices” $\Delta\omega_{m,n} = \{\omega : x_{m,n} \leq X(\omega) < x_{m,n+1}\}$ for $0 \leq n \leq m2^m - 1$ and $\Delta\omega_{m,m2^m} = \{\omega : X(\omega) \geq m\}$. In this notation we may write the simple approximation S_m to X given by (XIII.4.1) in the form

$$S_m = \sum_{n=0}^{m2^m} x_{m,n} \mathbf{1}_{\Delta\omega_{m,n}}.$$

As S_m increases pointwise to X , the monotone convergence theorem tells us that $\mathbf{E}(S_m) \rightarrow \mathbf{E}(X)$. On the other hand, additivity of expectation shows that

$$\mathbf{E}(S_m) = \sum_{n=0}^{m2^m} x_{m,n} \mathbf{P}(\Delta\omega_{m,n}).$$

The sum on the right-hand side has the appearance of an approximation to an integral (where the domain of integration is the sample space Ω) and, as custom has made so natural, we emphasise this discovery in notation by writing

$$\mathbf{E}(S_m) \rightarrow \int_{\Omega} X(\omega) \mathbf{P}(d\omega),$$

the integral form on the right signifying for us, as for Leibniz three centuries ago, merely a limiting sum of a particular type. We may hence write

$$\mathbf{E}(X) = \int_{\Omega} X(\omega) \mathbf{P}(d\omega)$$

and a universal integral representation for expectation has emerged.

More generally, if $X = X^+ - X^-$ is an integrable random variable its expectation, $\mathbf{E}(X)$, when it exists, can be expressed in this new notation as

$$\mathbf{E}(X) = \int_{\Omega} X^+(\omega) \mathbf{P}(d\omega) - \int_{\Omega} X^-(\omega) \mathbf{P}(d\omega).$$

We agree to compact notation and combine the two integral representations on the right and write the expectation of X simply as

$$\int_{\Omega} X(\omega) \mathbf{P}(d\omega) \quad \text{or, even more succinctly,} \quad \int_{\Omega} X d\mathbf{P}.$$

Too much should not be read into the infinitesimal appearance of the quantity $d\omega$ in these integrals. Instead the integrals should be taken at face value to mean the limit of an approximating sum, the notation reminiscent of the more ordinary kind of integral. It should be emphasised that this notation is a matter of *definition*—these integrals are just another, more complicated, way of writing $E(X)$ with the virtue, at least, of making explicit the probability measure at hand. While in principle one might imagine computing a Lebesgue integral, that is to say, expectation, by building up a sequence of approximating simple functions and then appealing to the monotone convergence theorem and taking limits, in practice one hardly ever follows such a route. There are usually far simpler ways of computing expectations directly via the distribution function of the random variable as we've seen in the discrete and absolutely continuous cases or, more frequently, by using additivity and other properties of expectation to reduce the problem to the consideration of simpler expectations.

If A is any event then $X1_A$ stands for the random variable that takes value $X(\omega)$ when ω is in A and is zero otherwise. We then write $\int_A X(\omega) P(d\omega)$ or $\int_A X dP$ for the expectation, $E(X1_A)$, of the truncated variable $X1_A$. In particular, $E(1_A) = P(A) = \int_A dP$ are all notational variants for the probability of the event A . Thus, as remarked earlier, *probabilities may be identified with the expectation of the corresponding indicator or, equivalently, an integral of the probability measure over the set in question.*

These integrals are called *Lebesgue integrals* after their founder Henri Lebesgue who introduced them in 1902 in a marvellously supple and powerful new theory of integration in his classic memoir *Intégrale, Longueur, Aire*.

The value of the integral notation is not that it exposes a hitherto unsuspected property of expectation—it doesn't—but that, by providing a spare and simple description of a limiting sum, it allows us to focus on essentials. The hard-won properties of additivity, homogeneity, and monotonicity of expectation ensure that these new integrals function just as we expect integrals to: they are additive, homogeneous, and monotone. In this light, we may interpret Theorem XIII.5.3 as a statement of the permissibility of (Lebesgue) integration of a series term by term. Calling these objects integrals now requires no great exercise of the imagination—if it walks like a duck and quacks like a duck it must be a duck.

THE VIEW FROM $(\mathbb{R}, \mathcal{B}, F)$

Another variant of the notation is suggested by the summands on the right of (XIII.4.4): the expectation of a positive random variable X is approximated by a sum of weighted differentials $\Delta F = F(2^{-m}n, 2^{-m}(n+1))$,

$$E(S_m) = \sum_{n=0}^{2^m-1} \frac{n}{2^m} F\left(\frac{n}{2^m}, \frac{n+1}{2^m}\right) + mF[m, \infty),$$

the last term on the right vanishing asymptotically if X has finite expectation.

The sum on the right has the unmistakable look of an approximation to an integral and, passing to the limit as $m \rightarrow \infty$, it is natural to write

$$E(S_m) \rightarrow \int_{\mathbb{R}^+} x F(dx),$$

the integral on the right standing for the expectation of X . In the general case the expectation $E(X) = E(X^+) - E(X^-)$, when it exists, of a real-valued random variable $X = X^+ - X^-$ may now be expressed in this notation as

$$E(X) = \int_{\mathbb{R}^+} x F^+(dx) - \int_{\mathbb{R}^+} x F^-(dx)$$

where F^+ and F^- denote the d.f.s of the positive and negative parts X^+ and X^- , respectively. As before, we combine the integral difference into a compact notation and write simply

$$\int_{\mathbb{R}} x F(dx) \quad \text{or, alternatively,} \quad \int_{\mathbb{R}} x dF(x)$$

for the expectation of X . The integral on the right appears, somewhat anachronistically from the measure-theoretic point of view, in terms of an integral involving the d.f. F viewed as a point function instead of as a set function as is implicit on the left, and is called a *Stieltjes integral*. The reader should interpret these expressions, however, as just other ways of writing the Lebesgue integral, that is to say, mathematical expectation, of an integrable random variable X . In a bow to history and convention, these integral forms, all standing for the same thing, are hence sometimes referred to as *Lebesgue–Stieltjes integrals*.

If \mathbb{I} represents any Borel set on the line, then $X1_{\mathbb{I}}(X)$ represents the truncated variable which is equal to X on those sample points ω for which $X(\omega)$ lies in \mathbb{I} and is zero otherwise. In the Stieltjes representation it is now natural to write the expectation of $X1_{\mathbb{I}}(X)$ in the evocative integral form $\int_{\mathbb{I}} x dF(x)$. Some care has to be exercised with the boundaries of integration, however, if the point function notation is used. The integrals $\int_{(a,b]} x F(dx)$ and $\int_{[a,b)} x F(dx)$ differ exactly in the amount $b[F(b) - F(b-)]$ which is non-zero if F has a jump at b . We hence have to distinguish between the intervals $(a, b]$ and $[a, b)$ and the classical notation, if used, must amend to

$$\int_a^{b+} x dF(x) \quad \text{and} \quad \int_a^{b-} x dF(x),$$

respectively, to clearly distinguish between these cases. (We need not be concerned with the lower limit of integration in these two cases as the right continuity of F makes the cases $[a, b]$ and $(a, b]$ equivalent to $(a, b]$ and $[a, b)$, respectively, and in both cases the lower limit of integration must needs be identified

with $a+$.) Similarly, if $a_n \uparrow a$, we identify

$$\int_{a-}^{b+} x dF(x) \text{ and } \int_{a-}^{b-} x dF(x)$$

with the limits of $\int_{(a_n, b]} x dF(x)$ and $\int_{(a_n, b)} x dF(x)$, respectively, as $n \rightarrow \infty$.

One may think of the equivalent integral formulations $\int_{\Omega} X(\omega) P(d\omega)$ and $\int_{\mathbb{R}} x F(dx)$ for the expectation of X as being effected by a change of variable from ω to x , or vice versa. Just as the use of a Jacobian accommodates the change in infinitesimal volumes when moving from one coordinate system to another in ordinary integration, so too, with volume replaced more generally by measure, does the replacement of P by F account honestly for the adaptation in probability measure in moving between the spaces Ω and \mathbb{R} .

So, we now have three equivalent notations,

$$E(X) = \int_{\Omega} X(\omega) P(d\omega) = \int_{\mathbb{R}} x dF(x), \quad (1.1)$$

all ultimately standing for the expectation of X . Which is to be preferred? The form $E(X)$ is spare and elegant and identifies the random variable at hand—the identification of the underlying probability space is implicit. The Stieltjes integral notation $\int_{\mathbb{R}} x dF(x)$ goes a little further and identifies the distribution function of the underlying variable. Finally, the most elaborate integral notation, either $\int_{\Omega} X(\omega) P(d\omega)$ or the more succinct $\int_{\Omega} X dP$, explicitly identifies the probability space over which the integral is defined. The choice of notation depends on context; they all come to the same thing in the end.

INTEGRALS WITH RESPECT TO GENERAL MEASURES

The key difference between the Lebesgue theory of integration and the earlier theory of Riemann (which as we recall is intuitive and serves all the needs of elementary calculus) is subtle but potent. In Riemann's approach, the interval of integration is partitioned into subintervals. In Lebesgue's approach the domain is partitioned instead into measurable sets. A graphical cartoon of the two contrasting approaches to integration is shown in Figure 1: the x -axis connotes an abstract sample space in the figure while the y -axis represents the familiar range of the function. While in the Riemann idea of integration one approximates the area under a curve by parallel rectangular segments obtained by partitioning the domain (the "x-axis") into subintervals, in the Lebesgue idea of integration one achieves the same effect by partitioning the range of the function (the "y-axis") into subintervals thence inducing a partition of the domain into measurable sets which are not necessarily conjoint or increasing in any way.

While at first glance the change from the Riemann to the Lebesgue system may appear superficial, nonetheless something deep is being wrought here. The Riemann integral in one dimension is deeply dependent on the topology of

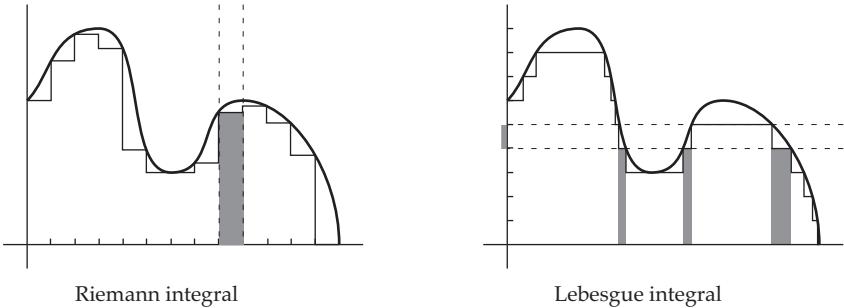


Figure 1: Partitioning the area under the curve.

the real line, in particular that it is an *ordered field*. The reader should consider how one would go about partitioning the domain if no such clear notion of order among the elements existed. (The Riemann integral in two dimensions doesn't really break new ground either as it is built up from the theory in one dimension by employing the Cartesian product of intervals in one dimension to partition the plane into two-dimensional rectangles; and similarly for higher dimensions.) In general, it is not clear how one would go about defining an intuitive notion of integration along these lines when the domain is an abstract space not equipped with a natural notion of order.

The Lebesgue theory of integration in sharp contradistinction relies upon the notion of order to partition the *range* of the function into small intervals. The domain can be arbitrary as long as it is equipped with a σ -algebra \mathcal{F} of measurable subsets of the space and a notion of size of those sets, that is to say, a positive measure $\mu: \mathcal{F} \rightarrow [0, \infty]$. In particular, it is not necessary that the space be equipped with a topological notion of order as for the real numbers nor is it necessary that the measure μ of the space be finite (as in the special case of probability measure which is our primary focus). The theory of integration can now be satisfactorily built up following the programme outlined in the previous chapter for any real- or complex-valued function X on such a space as long as it is measurable with respect to the underlying σ -algebra. Indeed, all that is required is to strike out stray references to \mathbf{P} in our development of expectation and replace them by μ . This permits us to extend and indeed build a theory of abstract integration on very general spaces and attach a natural meaning to the object $\int_{\Omega} X(\omega) \mu(d\omega)$ or, in short, $\int_{\Omega} X d\mu$, all the results of the previous chapter carrying over seamlessly to the framework of general Lebesgue integrals.

The case when the underlying space is a finite-dimensional Euclidean space is encountered frequently in applications. If Ω is the real line equipped with a measure μ on the Borel sets it is usual to write x instead of ω and, for any Baire function g on the real line, write the Lebesgue integral of g over the line in the form $\int_{\mathbb{R}} g(x) \mu(dx)$ or, simply, $\int_{\mathbb{R}} g d\mu$. For Baire functions and measures

on an n -dimensional Euclidean space \mathbb{R}^n equipped with measure μ , we write the sample points in vector notation as $x = (x_1, \dots, x_n)$ whence the integral becomes $\int \cdots \int_{\mathbb{R}^n} g(x_1, \dots, x_n) \mu(dx_1 \cdots dx_n)$ or $\int_{\mathbb{R}^n} g(x) \mu(dx)$ or $\int_{\mathbb{R}^n} g d\mu$, in increasing levels of notational compaction. If μ is Lebesgue measure λ then it is conventional to further shorten the description and write simply $\int_{\mathbb{R}} g(x) dx$ and $\int_{\mathbb{R}^n} g(x) dx$, respectively, in familiar integral notation.

As we have seen, the Lebesgue integral converges to familiar Riemannian forms when the latter is convergent. But the definition is much more flexible and general and permits consideration of a very wide range of functions that are not admissible under the Riemann theory. We are now in a position to close the loop in regard to the integrals we have encountered in sundry settings up to this point: *we henceforth agree to interpret integrals in the Lebesgue sense, the notion coinciding with the ordinary Riemann integrals in simple cases.*

From this vantage point expectations are special cases of Lebesgue integrals specialised to probability measures. Expectations in probabilistic settings, however, carry an intuitive sense of average or centre reinforced by our common experience. This aspect is somewhat obscured by the abstract formulation of integral. We will need very little of the theory of general Lebesgue integrals in this book excepting a couple of scattered instances where the application is so transparent as to need little comment.

2 Change of variable, moments, correlation

Expectation integrals take on familiar forms when the originating space is Euclidean in one or more coordinate variables. To avoid needless repetition we shall assume henceforth that we only deal with Baire functions and that equations involving expectations hold under the proviso that the expectations exist. Accordingly, suppose X is a random variable on some probability space and let F be its d.f. Suppose $W = g(X)$ is a Baire function and let G be its d.f. The random variable X now takes on the rôle of the coordinate variable (or sample point) on a real-line sample space equipped with measure F on the Borel sets, with g a measurable map on this space inducing a new probability space, $(\mathbb{R}, \mathcal{B}, F) \xrightarrow{g} (\mathbb{R}, \mathcal{B}, G)$. With the real line playing the rôle of the sample space Ω , we may hence, paralleling the alternative expressions in (1.1), write the expectation of W in any of the forms

$$E(W) = \int_{-\infty}^{\infty} g(x) dF(x) = \int_{-\infty}^{\infty} w dG(w).$$

This is sometimes given the grandiose title *fundamental theorem* though the two integral forms are obtained one from the other merely by the change of variable $w = g(x)$, the measures F and G accommodating accordingly.

The simplest of these settings occurs when $g(x) = x^n$ is a monomial.

DEFINITION 1 For any positive integer n , the quantity $E(X^n) = \int_{-\infty}^{\infty} x^n dF(x)$, if it exists, is called the *nth moment* of X . Thus, the first moment $E(X)$ is just the mean of the random variable X . The second moment is $E(X^2)$, and so on.

In applications we are frequently interested in quantifying how “concentrated” a random variable is about its mean. Suppose X has mean $m = E(X)$. The shift $W = X - m$ has the effect of “centring” the random variable X and by additivity it is clear that W has mean zero. The moments of such centred random variables convey information about how much the distribution spreads out from the mean.

DEFINITION 2 If X has mean m , the quantity $E((X - m)^n)$, if it exists, is called the *nth central moment of X* .

The second central moment $E\{(X - m)^2\}$ is the *variance* of X and denoted $\text{Var}(X)$; the positive square-root of the variance is the *standard deviation*. As we have seen, the variance carries a particular importance as a measure of the squared spread of the random variable away from its mean: a large variance implies that the random variable exhibits significant excursions from its mean value; a small variance on the other hand implies that the random variable is concentrated around its mean. If the reader is curious about what this means mathematically she will find a quantification via Chebyshev’s inequality in Section XVI.1.

The variance is intimately connected to the second moment. Indeed, by invoking linearity of expectation we obtain

$$\text{Var}(X) = E(X^2 - 2mX + m^2) = E(X^2) - 2mE(X) + m^2 = E(X^2) - m^2,$$

or, equivalently, $E(X^2) = \text{Var}(X) + E(X)^2$. As the expectation of a positive random variable is positive it follows that the variance is always positive. In particular, it follows that $E(X^2) \geq E(X)^2$ for any random variable whose first two moments exist. The reader has seen calculations of the mean and variance for several of the distributions of importance in Chapters VIII and IX.

Nothing much changes, excepting only a slightly more cumbersome notation, if the univariate Baire function is replaced by a Baire function of two or more variables. For instance, suppose $W = g(X_1, \dots, X_n)$ is a Baire function of the coordinate variables X_1, \dots, X_n . If (X_1, \dots, X_n) has d.f. $F(x_1, \dots, x_n)$ and W has d.f. $G(y)$, then

$$E(W) = \int \cdots \int_{\mathbb{R}^n} g(x_1, \dots, x_n) dF(x_1, \dots, x_n) = \int_{\mathbb{R}} w dG(w).$$

While at some level, the fundamental theorem is just a statement of a change of variables formula for integration, it packs a useful punch in practice. *The mean of a transformed random variable $W = g(X_1, \dots, X_n)$ is expressible directly in terms of a Lebesgue–Stieltjes integral involving the distribution of the original random variables*

X_1, \dots, X_n only. In particular, it is unnecessary to go through a preliminary computation to explicitly determine the distribution of W before computing its mean. The following example illustrates the principle.

EXAMPLE 1) *Waveforms with random phase; stationarity.* Sinusoidal waveforms form the backbone of many electrical communication systems (in which contexts they are usually called “carriers”). Consider the transmission of a pure tone at radian frequency ν with a random delay or phase Θ . The waveform then has the form $X_t = A \cos(\nu t + \Theta)$ where Θ is a random phase factor. Suppose Θ is uniform over $[0, 2\pi]$. Then, for any t , an application of the fundamental theorem shows

$$E(X_t) = \frac{1}{2\pi} \int_0^{2\pi} A \cos(\nu t + \theta) d\theta = 0.$$

A very reasonable result as the observed value of the waveform at any instant of time t has a symmetric distribution around positive and negative values.

The value of X_t at any given point in time t influences its value at other points. Let t and $t + \tau$ be two points in time separated by a given amount τ and consider the random variable $W_{t,t+\tau} = X_t X_{t+\tau}$. The expected value of $W_{t,t+\tau}$ then constitutes a crude measure of the level of dependence in the waveform values at times t and $t + \tau$. Another application of the change of variable formula for expectation shows that

$$\begin{aligned} E(W_{t,t+\tau}) &= \frac{A^2}{2\pi} \int_0^{2\pi} \cos(\nu t + \theta) \cos(\nu t + \nu\tau + \theta) d\theta \\ &= \frac{A^2}{4\pi} \int_0^{2\pi} \cos(2\nu t + \nu\tau + 2\theta) d\theta + \frac{A^2}{4\pi} \int_0^{2\pi} \cos(\nu\tau) d\theta = \frac{A^2}{2} \cos(\nu\tau). \end{aligned}$$

Thus, the expectation of $W_{t,t+\tau}$ depends only on the separation τ between the two points t and $t + \tau$ and is independent of t . The reader is encouraged to try to show indeed that the pair X_t and $X_{t+\tau}$ has a joint distribution determined solely by τ and, more generally, for any positive integer n , the joint distribution of X_{t_1}, \dots, X_{t_n} is determined by the time differences $t_i - t_j$ only. In other words, the finite-dimensional distributions of the random process X_t are invariant with respect to choice of time origin. Processes of this type are called *stationary*. ►

The expectation of a product of random variables has the form of a *mixed moment* and, as alluded to in the previous example, connotes a measure of the dependency of the two variables.

DEFINITION 3 Suppose the random pair (X, Y) has d.f. $F(x, y)$. Then the expectation of the product of the pair, $E(XY) = \iint_{\mathbb{R}^2} xy dF(x, y)$, is called the *correlation* between X and Y (with the usual proviso on existence). The centred mixed moment $E\{(X - E(X))(Y - E(Y))\}$ is called the *covariance* between X and Y and denoted $Cov(X, Y)$.

As an easy application of additivity of expectation the reader should verify that $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$. If X is identified with Y then the expressions for the covariance reduce to that of the variance.

The correlation and covariance carry (crude) information about the dependency structure of the random variables. A scale-invariant measure may be obtained by normalisation with respect to the individual standard deviations. The *correlation coefficient* of X and Y is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}.$$

It is easy to see now that a shift and scale of X and Y does not change the correlation coefficient and, in particular, $\rho(X, Y)$ may be identified with the covariance of the zero-mean, unit-variance standardised variables $X^* = (X - E(X)) / \sqrt{\text{Var}(X)}$ and $Y^* = (Y - E(Y)) / \sqrt{\text{Var}(Y)}$.

We say that the random variables X and Y are *uncorrelated* if $E(XY) = E(X)E(Y)$ or, equivalently, $\text{Cov}(X, Y) = 0$. The link between correlation and independence is provided by the following basic result.

THEOREM 1 (PRODUCT RULE FOR EXPECTATION) *Suppose X and Y are independent random variables with finite expectation. Then X and Y are uncorrelated.*

Fubini's theorem in Section 5 suggests a simple proof by separation of variables in integration. The following proof is provided for the reader who would like to construct a proof from first principles.

PROOF: The waltz is always the same. We first establish the result for simple random variables, proceed to positive random variables courtesy our friendly neighbourhood monotone convergence theorem via a limiting sequence of simple random variables, and finish up for general random variables.

We recall that X and Y are independent random variables if, and only if, the σ -algebras generated by X and Y are independent, that is, every pair of events $A \in \sigma(X)$ and $B \in \sigma(Y)$ is independent.

Now suppose X and Y are independent simple random variables with representations $X(\omega) = \sum_{i=1}^n x_i 1_{A_i}(\omega)$ and $Y(\omega) = \sum_{j=1}^m y_j 1_{B_j}(\omega)$. The random variable XY then has representation $XY = \sum_{i=1}^n \sum_{j=1}^m x_i y_j 1_{A_i \cap B_j}$ where the events A_i and B_j are pairwise independent. It follows that

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j P(A_i \cap B_j) = \sum_i \sum_j x_i y_j P(A_i) P(B_j) \\ &= \left(\sum_i x_i P(A_i) \right) \left(\sum_j y_j P(B_j) \right) = E(X) E(Y). \end{aligned}$$

We next suppose X and Y are positive and independent. Then, as in (XIII.4.4), we may construct an increasing sequence of simple functions S_m which converges pointwise to X . Likewise, we can construct a corresponding sequence of increasing simple functions, say T_m , converging pointwise to Y . For each m , S_m has a representation in

terms of a finite measurable partition $\{A_{mn}, 0 \leq n \leq m2^m\}$ where each A_{mn} is an element of $\sigma(X)$; likewise, T_m has a representation in terms of a finite measurable partition $\{B_{mn}, 0 \leq n \leq m2^m\}$ where each B_{mn} is an element of $\sigma(Y)$. It follows that S_m and T_m are independent simple random variables whence $E(S_m T_m) = E(S_m) E(T_m)$. On the other hand, $0 \leq S_m \uparrow X, 0 \leq T_m \uparrow Y$, and $0 \leq S_m T_m \uparrow XY$, so that we conclude via the monotone convergence theorem that $E(XY) = E(X) E(Y)$.

Finally, suppose X and Y are independent. By a decomposition of $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ into the constituent positive and negative parts, we see that $XY = X^+Y^+ - X^+Y^- - X^-Y^+ + X^-Y^-$. Taking expectations of both sides shows hence that

$$\begin{aligned} E(XY) &= E(X^+Y^+) - E(X^+Y^-) - E(X^-Y^+) + E(X^-Y^-) \\ &= E(X^+)E(Y^+) - E(X^+)E(Y^-) - E(X^-)E(Y^+) + E(X^-)E(Y^-). \end{aligned}$$

The first step follows by additivity of expectation, the second because the maps $X \mapsto (X^+, X^-)$ and $Y \mapsto (Y^+, Y^-)$ result in independent pairs of positive variables (by Theorem XII.8.2', for instance). By collecting terms on the right we see hence that

$$E(XY) = [E(X^+) - E(X^-)][E(Y^+) - E(Y^-)] = E(X)E(Y)$$

and so X and Y are uncorrelated. ▶

The result can be extended to a slightly more general product property for expectation.

THEOREM 2 *Let X_1, \dots, X_n be independent random variables with finite expectation. Then*

$$E(X_{i_1} X_{i_2} \cdots X_{i_k}) = E(X_{i_1}) E(X_{i_2}) \cdots E(X_{i_k}) \quad (2.1)$$

for every $1 \leq k \leq n$ and every subcollection of k out of the n random variables.

We've already seen that the touted product property holds for $k = 2$ in Theorem 1. An easy induction finishes off the proof.

There is sometimes a temptation to equate uncorrelatedness with independence on vague grounds; the assertion is not merely sloppy, in general it is wrong: *if X and Y are uncorrelated they are not, in general, independent*. Two counterexamples serve to reinforce this point.

EXAMPLES: 2) Suppose X takes values in ± 1 and ± 2 only, each with probability $1/4$. Let $Y = X^2$. Then $E(XY) = E(X^3) = \frac{1}{4}(-8 - 1 + 1 + 8) = 0$. On the other hand, $E(X)$ is clearly 0 so that $E(XY) = E(X)E(Y) = 0$. Thus, X and Y are uncorrelated. They are manifestly dependent, however; in fact, Y is completely determined by X .

3) Suppose U and V are independent random variables with a common distribution. For definiteness, suppose that they are absolutely continuous with a common d.f. G and corresponding density g . Form the random variables $X = U + V$ and $Y = U - V$. Then $E(XY) = E(U^2 - V^2) = E(U^2) - E(V^2) = 0$. On

the other hand, $E(Y) = E(U - V) = E(U) - E(V) = 0$. Thus, $E(XY) = E(X)E(Y)$ and the random variables X and Y are uncorrelated.

Now, the occurrence of the event $\{X \leq x, Y \leq y\}$ implies, and is implied by, the system of inequalities $U \leq (x + y)/2$, $V \leq x - U$, and $V \geq U - y$. Integrating out the density $g(u)g(v)$ of the pair (U, V) over the region indicated by these inequalities we see that the joint d.f. $F(x, y)$ of (X, Y) is given by

$$\begin{aligned} F(x, y) &= \int_{-\infty}^{(x+y)/2} \int_{-\infty}^{x-u} g(u)g(v) dv du - \int_{-\infty}^{(x+y)/2} \int_{-\infty}^{u-y} g(u)g(v) dv du \\ &= \int_{-\infty}^{(x+y)/2} G(x-u)g(u) du - \int_{-\infty}^{(x+y)/2} G(u-y)g(u) du. \end{aligned}$$

A similar, but easier, argument shows that the d.f., F_1 , of X is given by

$$F_1(x) = \int_{-\infty}^{\infty} G(x-u)g(u) du,$$

while, likewise, the d.f., F_2 , of Y is given by

$$F_2(y) = \int_{-\infty}^{\infty} [1 - G(u-y)]g(u) du = 1 - \int_{-\infty}^{\infty} G(u-y)g(u) du.$$

It is easy to verify that in general $F(x, y) \neq F_1(x)F_2(y)$. (The reader may consider the case where U and V are uniform over the unit interval.) Thus, X and Y are uncorrelated but not, in general, independent. ▶

In view of these examples the fact seen in Example VII.5.5 that uncorrelated (jointly) normal variables are independent exposes once more the special nature of the normal. In order to draw this conclusion, however, it is important that the variables in question are *jointly* normal; it is not sufficient if the variables X and Y are merely *marginally* normal. As a cautionary tale I provide

EXAMPLE 4) *Uncorrelated marginal normals need not be independent.*¹ Let X be a standard normal random variable with mean 0 and variance 1 and suppose ζ is any positive real number. Let $Y = -X \cdot 1(|X| \leq \zeta) + X \cdot 1(|X| > \zeta)$ be the random variable taking value $-X$ if $|X| \leq \zeta$ and value X if $|X| > \zeta$. As the density $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ of X is an even function it follows readily that the random variable Y is also distributed according to a standard normal distribution with mean 0 and variance 1. Thus, both X and Y have marginal $N(0, 1)$ distributions. As $XY = -X^2$ when $|X| \leq \zeta$, and $XY = X^2$ when $|X| > \zeta$, it follows that

$$E(XY) = \int_{-\infty}^{-\zeta} x^2 \phi(x) dx - \int_{-\zeta}^{\zeta} x^2 \phi(x) dx + \int_{\zeta}^{\infty} x^2 \phi(x) dx = 4 \int_{\zeta}^{\infty} x^2 \phi(x) dx - 1.$$

¹I am indebted to Jan Eriksson for bringing this variant on a classical theme to my attention.

The right-hand side is a continuous function, say $h(\zeta)$, of the positive parameter ζ . As $\int_{-\infty}^{\infty} x^2 \phi(x) dx = 1$ (this is the variance of the normal), it follows moreover that $h(\zeta)$ decreases monotonically from a value of +1 at $\zeta = 0$ to a limiting value of -1 as ζ increases to $+\infty$. It follows that the graph of h has a zero crossing, $h(\zeta) = 0$, at some point $\zeta = \zeta^*$. (This is the observation from elementary calculus known as Bolzano's theorem; a numerical evaluation shows $\zeta^* \approx 1.53817$ but the exact value is not germane for our purposes.) With this choice of $\zeta = \zeta^*$ it follows that $E(XY) = 0$. As X has zero mean it follows that $E(XY) = E(X)E(Y)$. The random variables X and Y are hence uncorrelated. On the other hand, Y is completely determined by X so that they are patently dependent. Thus, *uncorrelated marginally normal random variables are not necessarily independent*. The conclusion of Example VII.5.5 stands in contrast: joint normality is a much more demanding attribute of distributions than marginal normality. ►

In our consideration of the elementary arithmetic and continuous distributions we discovered that another additive property—for the variance this time—appears when we form sums of independent variables. Uncorrelatedness provides the link.

THEOREM 3 (ADDITIVITY OF VARIANCE) Suppose X_1, \dots, X_n are pairwise uncorrelated, i.e., $E(X_i X_j) = E(X_i)E(X_j)$ for all $i \neq j$, and let $S_n = X_1 + \dots + X_n$. Then

$$\text{Var}(S_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n).$$

The result holds a fortiori if the random variables $\{X_i\}$ are independent, or even if they are just pairwise independent.

PROOF: The proof for $n = 2$ is simple and provides a model for a general inductive proof. Writing $\mu_1 = E(X_1)$ and $\mu_2 = E(X_2)$, by additivity, $(X_1 + X_2) - E(X_1 + X_2) = (X_1 - \mu_1) + (X_2 - \mu_2)$. By squaring both sides and taking expectations, another application of additivity shows that

$$\begin{aligned} \text{Var}(X_1 + X_2) &= E[(X_1 - \mu_1)^2] + E[(X_2 - \mu_2)^2] + 2E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \text{Var}(X_1) + \text{Var}(X_2) + 2\text{Cov}(X_1, X_2). \end{aligned}$$

The final term on the right vanishes as X_1 and X_2 are uncorrelated by hypothesis and thus $\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$. The general result follows by repeated application of this result. By additivity of expectation again,

$$\begin{aligned} E(S_{n-1}X_n) &= E(X_1 X_n) + \dots + E(X_{n-1} X_n) \\ &= E(X_1)E(X_n) + \dots + E(X_{n-1})E(X_n) = E(S_{n-1})E(X_n) \end{aligned}$$

as X_n is uncorrelated with each of X_1, \dots, X_{n-1} , and so S_{n-1} and X_n are uncorrelated. As $S_n = S_{n-1} + X_n$, by the just-proved result it follows that $\text{Var}(S_n) = \text{Var}(S_{n-1}) + \text{Var}(X_n)$ and an easy induction completes the proof. ►

The tantalising linearity of variance for the binomial, negative binomial, and gamma distributions is now explicable. While not of the same order of surpassing generality as the addition theorem for expectation, the addition theorem for variances crops up often enough in applications to make it worth remembering.

3 Inequalities via convexity

The theory of inequalities in mathematics forms a rich and glorious tapestry, the classical inequalities taking on a particularly beguiling and intuitive form in probabilistic settings. These will be of repeated use in applications. We begin our investigation of these topics by borrowing a soupçon from the canon.

Convex functions are a staple of analysis, the notion of convexity key to many of the fundamental inequalities. We begin promptly with a definition. A *convex combination* of points x and y on the real line is an expression of the form $t = \alpha x + (1 - \alpha)y$ where $0 \leq \alpha \leq 1$. As α varies from 0 to 1, the convex combination t ranges through all points in the closed interval with endpoints x and y .

DEFINITION A real-valued function ψ on an open interval \mathbb{I} (which may be finite or infinite) is *convex*² if it satisfies the inequality

$$\psi(\alpha x + (1 - \alpha)y) \leq \alpha\psi(x) + (1 - \alpha)\psi(y) \quad (3.1)$$

for every convex combination of points in \mathbb{I} . We say that ψ is *strictly convex* if the inequality in (3.1) is strict whenever $x \neq y$ and $\alpha \notin \{0, 1\}$.

Geometrically, (3.1) says that the chord XY connecting the points $X = (x, \psi(x))$ and $Y = (y, \psi(y))$ in the plane lies above the graph of the function in the interval from x to y . Our definition is equivalent to the statement that the inequality

$$\frac{\psi(t) - \psi(s)}{t - s} \leq \frac{\psi(u) - \psi(t)}{u - t} \quad (3.1')$$

holds whenever $s < t < u$. Indeed, if the inequality were violated for some t then the point $T = (t, \psi(t))$ in the plane would lie above the chord SU connecting the points $S = (s, \psi(s))$ and $U = (u, \psi(u))$ so that the graph of the function would lie above the chord SU at least at the point with abscissa t . [This is geometrically clear but the reader of a more particular disposition who wishes to verify this algebraically should consider the consequences of supposing that (3.1') is violated at some interior point t of the open interval (s, u) .]

²We also use this terminology if ψ is a real-valued function on \mathbb{R}^n satisfying (3.1). In this case we should naturally interpret the sum in the argument of ψ on the left as a vector-valued convex combination.

This would then imply that $\psi(t) > \frac{u-t}{u-s}\psi(s) + (1 - \frac{u-t}{u-s})\psi(u)$. Identifying $\alpha = (u-t)/(u-s)$ this leads, on the one hand, to the convex combination $t = \alpha s + (1-\alpha)u$ and, on the other, to the inequality $\psi(t) > \alpha\psi(s) + (1-\alpha)\psi(u)$. Contradiction.] The inequality (3.1') says very simply that the slope of any chord to the right of any given point exceeds the slope of any chord to the left of that point. If the reader now sketches a few curves she will see that a convex function has a cup-like character and a little experimentation will suggest to her that the graph cannot have any breaks.

THEOREM 1 *If ψ is convex then it is continuous and a fortiori Borel measurable.*

PROOF: It is simplest to appeal to the geometry of the situation. Suppose $u < x < y < v$ are points on the abscissa and let U, X, Y , and V be the corresponding points on the graph of the function. Here $U = (u, \psi(u))$ and so on. As shown in Figure 2, Y must lie above the line UX and below the line XV . [To assume the contrary, suppose, for instance, that Y lies below the line UX . As x is an interior point of the interval (u, y) , we are led to the conclusion that the point X lies above the chord UY . But this contradicts (3.1).] Let $\underline{Y} = (y, \underline{\psi}(y))$ and $\bar{Y} = (\bar{y}, \bar{\psi}(y))$ be the points corresponding to abscissa y on the lines UX and XV , respectively. Then Y is sandwiched between \underline{Y} and \bar{Y} ; equivalently, $\underline{\psi}(y) < \psi(y) < \bar{\psi}(y)$. But the points \underline{Y} and \bar{Y} may be brought as close to X , and hence to each other, as desired by moving y close to x . Formally, for any $\epsilon > 0$ we may select $\delta > 0$ such that if $0 < y - x < \delta$ then $|\psi(y) - \psi(x)| < \epsilon$ and $|\bar{\psi}(y) - \psi(x)| < \epsilon$. (Linear forms are continuous!) But then this implies that $|\psi(y) - \psi(x)| < \epsilon$ and so ψ is continuous from the right at x . A similar argument shows that ψ is continuous from the left at x also. ▶

If, additionally, ψ is differentiable then, by the formulation (3.1'), it is clear that ψ' must be an increasing function. And if ψ is fortunate enough to have two derivatives then it must be the case that $\psi''(x) \geq 0$ for each x . This provides a simple test for convexity.

THEOREM 2 *If ψ is twice differentiable and $\psi''(x) \geq 0$ for all x then ψ is convex.*

Thus, x^2 and e^x are convex on $(-\infty, \infty)$ and $-\log x$ is convex on the open half-line $(0, \infty)$.

A very profitable line of thought is opened up by the observation that the convex combinations on both sides of (3.1) may be interpreted as expectations with respect to the distribution which places mass α at x and mass $1 - \alpha$ at y . The wonderfully flexible inequality of Jensen says that the averaging implicit in these expressions may be extended to arbitrary expectations.

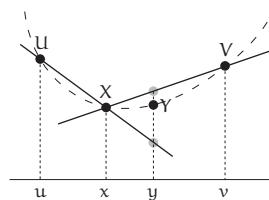


Figure 2: Convexity implies continuity.

JENSEN'S INEQUALITY *Suppose ψ is a convex function on the real line, X an integrable random variable. If $\psi(X)$ is integrable then $\psi(E(X)) \leq E(\psi(X))$. Equality holds when ψ is strictly convex if, and only if, X has a degenerate distribution concentrated on one point.*

PROOF: Set $t = E(X)$ and by letting s vary in the left-hand side of (3.1'), let $\beta(t) = \sup\{\frac{\psi(t)-\psi(s)}{t-s} : s < t\}$. By definition of supremum, $\beta(t)$ is a lower bound for the right-hand side of (3.1') for every $u > t$; on the other hand it is also manifestly an upper bound for the left-hand side of (3.1') for every $s < t$. It follows that

$$\frac{\psi(t)-\psi(s)}{t-s} \leq \beta(t) \leq \frac{\psi(u)-\psi(t)}{u-t}$$

for every $s < t < u$. Isolating $\psi(s)$ by a consideration of the first inequality and $\psi(u)$ by a consideration of the second inequality it becomes clear that the inequality $\psi(x) \geq \psi(t) + \beta(t)(x - t)$ holds for every x . Replacing x by the random variable X we obtain $\psi(X) \geq \psi(t) + \beta(t)(X - t)$. Taking expectations of both sides and bearing in mind that $E(X) = t$ we obtain $E(\psi(X)) \geq \psi(E(X))$ by monotonicity and linearity of expectation. To complete the proof, as the inequality (3.1') is strict if ψ is strictly convex, it follows that $\psi(x) > \psi(t) + \beta(t)(x - t)$ if $x \neq t$ whence $\psi(x) = \psi(t)$ if, and only if, $x = t$. Thus, $\psi(E(X)) = E(\psi(X))$ if, and only if, X is concentrated at the point t . ▶

We may derive a variety of familiar, and not so familiar, inequalities by selecting for ψ various elementary functions.

A QUADRATIC INEQUALITY: The selection of quadratic $\psi(x) = x^2$ recovers the inequality $E(X)^2 \leq E(X^2)$ which we had deduced from elementary considerations as $0 \leq \text{Var}(X) = E(X^2) - E(X)^2$. The logarithm and exponential functions yield more interesting inequalities.

A LOGARITHMIC INEQUALITY: Suppose $P = \{p_k\}$ and $Q = \{q_k\}$ are two discrete distributions indexed, say, by the integers. The *Kullback–Leibler divergence between P and Q* is defined to be the extended real-valued function

$$D(P, Q) = \sum_k p_k \log\left(\frac{p_k}{q_k}\right). \quad (3.2)$$

[We adopt the natural convention $0 \log(0) = 0$.] The utility of this function in comparing distributions rests upon the following result.

THEOREM 4 *Let P and Q be discrete distributions. Then $D(P, Q) \geq 0$ with equality if, and only if, $P = Q$.*

PROOF: If we introduce a discrete random variable X which takes value q_k/p_k with probability p_k then we may identify $D(P, Q) = E(-\log(X))$. Now $E(X) =$

$\sum_k p_k \cdot \frac{q_k}{p_k} = \sum_k q_k = 1$. As $\psi(x) = -\log(x)$ is convex on $(0, \infty)$ the stage is set for an application of Jensen's inequality and we obtain

$$D(P, Q) = E(-\log(X)) \geq -\log(E(X)) = -\log 1 = 0.$$

As the second derivative of $-\log(x)$ is strictly positive, the inequality in (3.1) is strict unless $x = y$. This implies that equality can be obtained in Jensen's inequality if, and only if, X has a degenerate distribution concentrated on one point. But this means that $D(P, Q) = 0$ if, and only if, $p_k = q_k$ for each k . ►

The Kullback–Leibler divergence is in wide use in analysis as it provides a tractable notion of “distance” in the space of distributions. In Shannon’s theory of information it crops up under the nom de plume *relative entropy*. While the divergence shares some of the characteristic features of distance, positivity, and uniqueness of zero, it is not a genuine metric because it is not symmetric: $D(P, Q) \neq D(Q, P)$ in general.

AN EXPONENTIAL INEQUALITY: A consideration of the exponential function yields another class of familiar and important inequalities. Let Y be a random variable taking values in a countable collection of points $\{y_k\}$ on the line, $\{p_k\}$ its probability distribution which assigns mass p_k to the point y_k . Jensen's inequality applied to the function $\psi(x) = e^x$ then says that

$$\prod_k e^{p_k y_k} = e^{E(Y)} \leq E(e^Y) = \sum_k p_k e^{y_k}. \quad (3.3)$$

Identifying $x_k = e^{y_k}$ allows us to write the inequality in a compact form.

THE INEQUALITY OF ARITHMETIC AND GEOMETRIC MEANS *Let $\{p_k\}$ be a discrete probability distribution, $p_k > 0$ for each k , and $\{x_k\}$ any countable collection of points in the positive half-line \mathbb{R}^+ . Then $\prod_k x_k^{p_k} \leq \sum_k p_k x_k$. Equality holds if, and only if, there exists a constant x such that $x_k = x$ for each k .*

It only remains to verify the condition for equality. The bound in (3.3) holds with equality if, and only if, Y is concentrated at one point, and as all the probabilities p_k are strictly positive this can only occur if each of the y_k 's is equal to some common value $y = \log(x)$.

In the case where the distribution places equal mass $1/n$ on each of the points $y_1 = \log x_1, \dots, y_n = \log x_n$, we obtain the familiar inequality

$$(x_1 x_2 \cdots x_n)^{1/n} \leq \frac{1}{n}(x_1 + x_2 + \cdots + x_n) \quad (3.4)$$

which says that the geometric mean of a sequence is bounded above by its arithmetic mean. If we consider a parallelepiped with sides x_1, x_2, \dots, x_n in \mathbb{R}^n then $V = x_1 x_2 \cdots x_n$ represents its volume and $S = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ its average side length. Then (3.4) may be cast in the form $V \leq S^n$ and the inequality of arithmetic and geometric means adapted to this situation says that among

all parallelepipeds of a given mean side length S , the cube of side S has the largest volume. This is the “rectangular version” of the classical isoperimetric inequality that the reader may be familiar with (see Section 8).

J. M. Steele has pointed out that the pervasive utility of the inequality of arithmetic and geometric means may be tied to its origins in the exponential function $x \mapsto e^x$ which plays a fundamental rôle in linking two fundamental groups in mathematics—the reals equipped with addition and the positive reals equipped with multiplication. The very useful inequality attributed to Hölder provides a case in point in the following section.

4 L^p -spaces

We consider random variables on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For each $p \geq 1$, the space $L^p = L^p(\Omega, \mathcal{F}, \mathbf{P})$ denotes the family of random variables X for which $\mathbf{E}(|X|^p)$ is finite. In this terminology, L^1 is the space of integrable random variables, L^2 the space of square-integrable random variables, and so on.

The introduction of a little notation helps lubricate the flow. Suppose $p \geq 1$ and X is in L^p . Then $\mathbf{E}(|X|^p)$ is finite and we write $\|X\|_p = \mathbf{E}(|X|^p)^{1/p}$. The quantity $\|X\|_p$ is positive and stands in the rôle of *norm* in the space L^p in a sense that we will make precise shortly. A preview of its normalisation attribute is seen in the consideration of the scaled variable $X/\|X\|_p$. Whenever $\|X\|_p \neq 0$ we have

$$\left\| \frac{X}{\|X\|_p} \right\|_p = \mathbf{E} \left(\frac{|X|^p}{\|X\|_p^p} \right)^{1/p} = \frac{1}{\|X\|_p} \mathbf{E}(|X|^p)^{1/p} = 1$$

and scaling by $\|X\|_p$ reduces non-zero variables to unit length.

Real numbers $p > 1$ and $q > 1$ are said to form a pair of *conjugate exponents* if $\frac{1}{p} + \frac{1}{q} = 1$. Thus, if p and q are conjugate exponents then $\{p^{-1}, q^{-1}\}$ forms a discrete (Bernoulli) distribution.³ Now suppose X and Y are positive random variables, $X \in L^p$ and $Y \in L^q$, for a pair of conjugate exponents p and q . The inequality of arithmetic and geometric means then says that $X^{1/p} \cdot Y^{1/q} \leq \frac{1}{p}X + \frac{1}{q}Y$. If X and Y are not concentrated at the origin then $\|X\|_p > 0$ and $\|Y\|_q > 0$ and we may apply the inequality of arithmetic and geometric means to the scaled variables $\tilde{X} = (X/\|X\|_p)^p$ and $\tilde{Y} = (Y/\|Y\|_q)^q$ to obtain

$$\frac{XY}{\|X\|_p \cdot \|Y\|_q} = \tilde{X}^{1/p} \cdot \tilde{Y}^{1/q} \leq \frac{1}{p}\tilde{X} + \frac{1}{q}\tilde{Y} = \frac{1}{p} \cdot \frac{X^p}{\|X\|_p^p} + \frac{1}{q} \cdot \frac{Y^q}{\|Y\|_q^q}.$$

Taking expectations of both sides, we hence obtain the elegant inequality

$$\frac{\mathbf{E}(XY)}{\|X\|_p \cdot \|Y\|_q} \leq \frac{1}{p} + \frac{1}{q} = 1. \quad (4.1)$$

³The notation is unfortunate in the usage of p^{-1} and not p to represent a Bernoulli probability. But the convention is immutable in this context by reasons of tradition.

We may extend the result to general random variables via the modulus inequality $|\mathbf{E}(XY)| \leq \mathbf{E}(|X| \cdot |Y|)$ which reduces consideration on the right to a pair of positive variables $|X| \in L^p$ and $|Y| \in L^q$. The condition for equality in the inequality of arithmetic and geometric means says that equality holds in (4.1) if, and only if, $\tilde{X}^{1/p} = \tilde{Y}^{1/q}$ a.e. or, equivalently, X^p and Y^q are related by a linear form a.e.

HÖLDER'S INEQUALITY Suppose $p, q > 1$ are a pair of conjugate exponents, X is in L^p , and Y is in L^q . Then $|\mathbf{E}(XY)| \leq \|X\|_p \cdot \|Y\|_q$. If the variables are non-zero then equality holds if, and only if, there is a constant λ such that $\lambda X^p = Y^q$ a.e.

The case when X or Y is concentrated at the origin is trivial. Else we may identify $\lambda = \|Y\|_q / \|X\|_p$. Identifying $X \leftarrow |X|$ and $Y = 1$, we obtain a simple and useful result.

COROLLARY 1 (MONOTONICITY OF L^p -NORM) If $p \geq 1$ and $X \in L^p$ then X is integrable and $\|X\|_1 \leq \|X\|_p$.

The most familiar and widely used instance of Hölder's inequality is the case of equal conjugate exponents, $p = q = 2$. In this case the inequality reduces to a familiar statement which bounds the correlation of two square-integrable variables by the square-root of the product of their second moments.

COROLLARY 2 (CAUCHY–SCHWARZ INEQUALITY) If X and Y are square-integrable then $|\mathbf{E}(XY)| \leq \sqrt{\mathbf{E}(X^2)} \cdot \sqrt{\mathbf{E}(Y^2)}$.

By centring X and Y we may write the Cauchy–Schwarz inequality in the equivalent form

$$|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}$$

which provides another proof that the correlation coefficient ρ of the bivariate normal $\phi(\cdot, \cdot; \rho)$ given in (VII.5.3) must be bounded in absolute value by 1.

The map $X \mapsto \|X\|_p$ defines a positive function on the space L^p . The map deals with scaling in a natural fashion and homogeneity of expectation shows that $\|\lambda X\|_p = |\lambda| \cdot \|X\|_p$ for all real λ . Furthermore, the map displays the unique signature of length.

MINKOWSKI'S INEQUALITY Suppose $p \geq 1$. Then $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ for any pair of random variables in L^p .

PROOF: The standard triangle inequality $|X + Y| \leq |X| + |Y|$ establishes the result for the case $p = 1$. Suppose accordingly that $p > 1$. The simplest proof is due to F. Riesz and leverages Hölder's inequality. We may set the stage for the inequality by factoring the p th power into a product of terms via

$$|X + Y|^p = |X + Y| \cdot |X + Y|^{p-1} \leq |X| \cdot |X + Y|^{p-1} + |Y| \cdot |X + Y|^{p-1}.$$

If p and q are conjugate exponents then $p - 1 = p/q$ so that, by taking expectations of both sides and appealing to monotonicity, we obtain

$$\begin{aligned} \mathbb{E}(|X + Y|^p) &\leq \mathbb{E}(|X| \cdot |X + Y|^{p/q}) + \mathbb{E}(|Y| \cdot |X + Y|^{p/q}) \\ &\leq \mathbb{E}(|X|^p)^{1/p} \mathbb{E}(|X + Y|^p)^{1/q} + \mathbb{E}(|Y|^p)^{1/p} \mathbb{E}(|X + Y|^p)^{1/q}, \end{aligned}$$

the final step by two applications of Hölder's inequality. The claimed result follows by factoring out the term $\mathbb{E}(|X + Y|^p)^{1/q}$ in the right-hand side. ►

The map $X \mapsto \|X\|_p$ hence satisfies the *triangle inequality* characteristic of any self-respecting measure of length or *norm*. If it walks like a duck and quacks like a duck it must be a duck. The space L^p is a *normed vector space*.

A careful reader may object to calling $\|\cdot\|_p$ a norm because the uniqueness of zero is not satisfied—if two random variables X and Y differ only on a set of measure zero then $\|X\|_p = \|Y\|_p$ and, in particular, $\|X\|_p = 0$ only guarantees that $X(\omega) = 0$ a.e. The standard way to bypass this objection is to identify variables which are equal almost everywhere. Say that X and Y are *equivalent* and write $X \sim Y$ if $X(\omega) = Y(\omega)$ a.e. The equivalence relation \sim partitions L^p into equivalence classes of variables, the variables in each of which are equivalent to each other; we say that L^p is *quotiented out by this equivalence relation*. This quotiented space inherits a bona fide norm from $\|\cdot\|_p$ —merely assign as norm to an equivalence class the value $\|X\|_p$ of any of its elements X . In probability theory, however, we hardly ever bother with such a fine distinction by quotienting; we think of L^p as a normed vector space simply bearing in mind that we identify elements that are equal a.e.



COMPLETENESS

It is a very important fact that every Cauchy sequence in the space L^p converges a.e., that is to say, the quotiented space is a *complete normed space* or *Banach space*. The method of proof will be familiar to the reader who knows that L^2 is complete.

THEOREM 3 *Every Cauchy sequence in L^p converges a.e. (to an element of L^p).*

PROOF: Suppose $\{X_n, n \geq 1\}$ is a Cauchy sequence in L^p , that is, $X_n \in L^p$ for each n and $\sup_{r,s \geq n} \|X_r - X_s\|_p \rightarrow 0$ as $n \rightarrow \infty$. Then there exists an increasing subsequence $\{n_k, k \geq 1\}$ such that $\sup_{r,s \geq n_k} \|X_r - X_s\|_p \leq 2^{-k}$ for each $k \geq 1$. In particular, by Corollary 1, $\mathbb{E}(|X_{n_{k+1}} - X_{n_k}|) = \|X_{n_{k+1}} - X_{n_k}\|_1 \leq \|X_{n_{k+1}} - X_{n_k}\|_p$, so that $\sum_{k=1}^{\infty} \mathbb{E}(|X_{n_{k+1}} - X_{n_k}|) \leq \sum_{k=1}^{\infty} 2^{-k} = 1$. By Theorem XIII.5.3, it follows that $X_{n_k} = X_{n_1} + \sum_{j=1}^{k-1} (X_{n_{j+1}} - X_{n_j})$ converges a.e. For definiteness, set $X(\omega) = \limsup_{k \rightarrow \infty} X_{n_k}(\omega)$. Then $X_{n_k}(\omega) \rightarrow X(\omega)$ a.e. We now have our candidate for limit. Our first order of business is to show that X is in L^p .

Fix any $j \geq 1$ and any $r \geq n_j$. Then, for any $k \geq j$, we have $\|X_r - X_{n_k}\|_p^p = \mathbb{E}(|X_r - X_{n_k}|^p) \leq 2^{-jp}$. By Fatou's lemma applied to the sequence $\{|X_r - X_{n_k}|^p, k \geq 1\}$,

$$\mathbb{E}(|X_r - X|^p) = \mathbb{E}(\liminf_{k \rightarrow \infty} |X_r - X_{n_k}|^p) \leq \liminf_{k \rightarrow \infty} \mathbb{E}(|X_r - X_{n_k}|^p) \leq 2^{-jp}, \quad (4.2)$$

and it follows that the difference $X_r - X$ is in L^p . As $X = X_r - (X_r - X)$ is the difference of two elements of L^p , it follows by Minkowski's inequality that X is in L^p as well. To finish off the proof, we observe from (4.2) that $\|X_r - X\|_p \leq 2^{-j}$ whenever $r \geq n_j$ and so $0 \leq \limsup_{r \rightarrow \infty} \|X_r - X\|_p \leq 2^{-j}$. As j may be taken arbitrarily large this implies that $\lim_{r \rightarrow \infty} \|X_r - X\|_p = 0$, whence $X_r(\omega) \rightarrow X(\omega)$ a.e. ▶

The case $p = 2$ is of special importance. The space L^2 (or, more precisely, L^2 quotiented out by equivalence) has the natural inner product $\langle X, Y \rangle = E(XY)$ which may be considered to induce the L^2 -norm $\|X\|_2 = \sqrt{\langle X, X \rangle} = \sqrt{E(X^2)}$. Thus, L^2 equipped with the inner product $\langle \cdot, \cdot \rangle$ is a *complete inner product space* or *Hilbert space*. The existence of an inner product imbues the space L^2 with a delicious geometric flavour, inner products connoting projections. In this regard, the completeness of L^2 is of particular importance as the geometric idea of projections onto subspaces can now be put on a firm footing.

If \mathcal{G} is a sub- σ -algebra of the parent σ -algebra \mathcal{F} then $L^2(\Omega, \mathcal{G}, P)$ represents the family of square-integrable, \mathcal{G} -measurable functions. As \mathcal{G} is coarser than \mathcal{F} , every \mathcal{G} -measurable function is naturally also \mathcal{F} -measurable though the converse is not true. In a natural sense the space $L^2(\Omega, \mathcal{G}, P)$ is a subspace of $L^2(\Omega, \mathcal{F}, P)$. Given any \mathcal{F} -measurable function X we may now consider its projection onto the subspace $L^2(\Omega, \mathcal{G}, P)$ as a candidate for “best approximation” by a \mathcal{G} -measurable function. This leads to an elegant geometric definition of conditional expectation (see Problems 46–51).

The inequalities of the last two sections are fundamental in analysis and the reader will find scattered applications through the rest of the book. If she casts her mind back across the train of inequalities she will find a salutary reminder of the importance of convexity as the wellspring from which these bounties flow. The reader interested in exploring further will find that age cannot wither, nor custom stale, Hardy, Littlewood, and Pólya’s magisterial treatise on the subject.⁴ Of more recent vintage—and a personal favourite—is J. M. Steele’s charming vignette.⁵

5 Iterated integrals, a cautionary example

Suppose a is any constant and $g(t) = t(a^2 + t^2)^{-1}$. Elementary differentiation with respect to t shows then that $g'(t) = (a^2 - t^2)/(a^2 + t^2)^2$. Accordingly,

$$\begin{aligned} \int_1^\infty \left(\int_1^\infty \frac{x^2 - y^2}{(x^2 + y^2)^2} dy \right) dx &= \int_1^\infty \frac{y}{x^2 + y^2} \Big|_{y=1}^\infty dx \\ &= - \int_1^\infty \frac{1}{1+x^2} dx = - \arctan(x) \Big|_1^\infty = -\frac{\pi}{2} + \frac{\pi}{4} = -\frac{\pi}{4} \end{aligned}$$

⁴G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Second Edition. Cambridge, UK: Cambridge University Press, 1962.

⁵J. M. Steele, *The Cauchy-Schwarz Masterclass*. Cambridge, UK: Cambridge University Press, 2004.

as we may treat x as constant in the inner integral. But, by interchanging the rôles of x and y , we obtain

$$\begin{aligned} \int_1^\infty \left(\int_1^\infty \frac{x^2 - y^2}{(x^2 + y^2)^2} dx \right) dy &= \int_1^\infty \frac{-x}{x^2 + y^2} \Big|_{x=1}^\infty dy \\ &= \int_1^\infty \frac{1}{1+y^2} dy = \arctan(y) \Big|_1^\infty = \frac{\pi}{2} - \frac{\pi}{4} = +\frac{\pi}{4}. \end{aligned}$$

This example is typical of the “proofs” one encounters that “ $1 = 0$ ”. The difficulty with the interchange in order of integration here is that the function $f(x, y) = (x^2 - y^2)/(x^2 + y^2)^2$ is not absolutely integrable over the region $[1, \infty] \times [1, \infty)$ in the plane. If absolute integrability can be assured then problems of this stripe do not occur and we may swap the order of integration with impunity. This basic but important result is called *Fubini’s theorem*.

Suppose $(\mathcal{X}, \mathcal{M}, \mu)$ and $(\mathcal{Y}, \mathcal{N}, \nu)$ are measure spaces where μ and ν are σ -finite measures on \mathcal{M} and \mathcal{N} , respectively. As sets of measure zero are irrelevant when dealing with integrals we may suppose, without loss of generality, that the σ -algebras \mathcal{M} and \mathcal{N} are complete (see Section XI.2).

Let $\mathcal{P} = \mathcal{M} \times \mathcal{N}$ denote the family of Cartesian products $A \times B$ with $A \in \mathcal{M}$ and $B \in \mathcal{N}$. As $(A_1 \times B_1) \cap (A_2 \times B_2) = (A_1 \cap A_2) \times (B_1 \cap B_2)$ it follows that \mathcal{P} is closed under intersections. In the language of Section III.5, \mathcal{P} is a π -class. In general the family \mathcal{P} is not in itself a σ -algebra so we enlarge it to equip the product space $\mathcal{X} \times \mathcal{Y}$ with the σ -algebra $\mathcal{M} \otimes \mathcal{N}$ generated by \mathcal{P} , in notation, $\mathcal{M} \otimes \mathcal{N} := \sigma(\mathcal{P})$. This is in much the same way as the Borel σ -algebra $\mathcal{B}(\mathbb{R}^2)$ on the plane is generated by the Cartesian product of intervals (or even Borel sets in the line). We are now equipped to consider the iterative integration of $\mathcal{M} \otimes \mathcal{N}$ -measurable functions on $\mathcal{X} \times \mathcal{Y}$. As usual, it is best to begin with a consideration of positive functions.

THEOREM 1 *Let f be any positive, extended real-valued function on $\mathcal{X} \times \mathcal{Y}$, measurable with respect to $\mathcal{M} \otimes \mathcal{N}$. Then the following three conditions hold: I1) the map $x \mapsto \int_{\mathcal{Y}} \nu(dy) f(x, y)$ is \mathcal{M} -measurable; I2) the map $y \mapsto \int_{\mathcal{X}} \mu(dx) f(x, y)$ is \mathcal{N} -measurable; and I3) $\int_{\mathcal{X}} \mu(dx) \int_{\mathcal{Y}} \nu(dy) f(x, y) = \int_{\mathcal{Y}} \nu(dy) \int_{\mathcal{X}} \mu(dx) f(x, y)$.*

This is the only preliminary result that needs technical finesse and the reader may wish to skip on by on a first reading to get to the *res gestae*.

The dance is always the same for the reader who wishes to construct a first-principles proof: begin with indicators for Cartesian products $A \times B$, progress to indicators of general measurable sets C in the product space, then to simple functions, and finally positive functions. The reader who attempts this will find that the only step in the programme outlined that is difficult is the second in the move from Cartesian products to general measurable sets (Problem XI.12 and Levi’s theorem of Section XIII.8 will help), but the procedure is undeniably tedious. The $\pi-\lambda$ theorem of Section III.5 allows

us to finesse the messy details and I shall do so here. A little terminology and notation will smooth the way.

We say that a sequence $\{f_n, n \geq 1\}$ of real-valued functions on some space Ω converges boundedly if it is uniformly bounded and converges pointwise. If \mathcal{A} is any family of subsets of Ω , we write $\chi(\mathcal{A})$ for the induced family of indicator functions 1_A as A sweeps over all the sets in \mathcal{A} and call $\chi(\mathcal{A})$ the *characteristic set* of \mathcal{A} . A recasting of the π - λ theorem of Section III.5 in functional terms helps the process along.

THE π - λ THEOREM, REVISITED Suppose \mathcal{H} is a linear subspace of bounded, real-valued functions on Ω which includes the constant functions and is closed under bounded convergence. If \mathcal{P} is any π -class whose characteristic set is contained in \mathcal{H} then $\chi(\sigma(\mathcal{P})) \subseteq \mathcal{H}$.

PROOF: The conditions on \mathcal{H} say that: (1) $1_\Omega \in \mathcal{H}$; (2) if $f, g \in \mathcal{H}$ then $f + g \in \mathcal{H}$; (3) if $f \in \mathcal{H}$ and $c \in \mathbb{R}$ then $cf \in \mathcal{H}$; and (4) if $f_n \in \mathcal{H}$ for each $n \geq 1$ and $f_n \rightarrow f$ boundedly then $f \in \mathcal{H}$. Now let \mathcal{L} be the family of subsets A of Ω for which 1_A is in \mathcal{H} , i.e., \mathcal{L} is the largest family of subsets of Ω whose characteristic set is contained in \mathcal{H} . Clearly then $\mathcal{P} \subseteq \mathcal{L}$ and by the π - λ theorem it suffices to establish that \mathcal{L} is a λ -class.

To begin, $1_\Omega \in \mathcal{H}$ and so $\Omega \in \mathcal{L}$. Now suppose $\{A_n, n \geq 1\}$ is an increasing family of sets in \mathcal{L} , $A_n \uparrow A = \bigcup_n A_n$. Then $1_{A_n} \in \mathcal{H}$ for each n and as it is clear that $1_{A_n}(\omega) \rightarrow 1_A(\omega)$ boundedly, we have $1_A \in \mathcal{H}$, whence $A \in \mathcal{L}$. Finally, suppose A, B are sets in \mathcal{L} with $A \subseteq B$. Then $1_A \in \mathcal{H}$ and $1_B \in \mathcal{H}$, whence $1_{B \setminus A} = 1_B - 1_A \in \mathcal{H}$, and so $B \setminus A \in \mathcal{L}$. Thus, \mathcal{L} is indeed a λ -class. ▶

PROOF OF THEOREM 1: We begin with the case of finite measures μ and ν , $\mu(\mathcal{X}) < \infty$ and $\nu(\mathcal{Y}) < \infty$. Introduce the linear subspace \mathcal{H} of bounded functions f on $\mathcal{X} \times \mathcal{Y}$, measurable with respect to $\mathcal{M} \otimes \mathcal{N}$, and which satisfy the integrability conditions I1), I2), and I3) given in the theorem. It is clear that $1_{\mathcal{X} \times \mathcal{Y}}(x, y) = 1_{\mathcal{X}}(x)1_{\mathcal{Y}}(y)$ is in \mathcal{H} and so \mathcal{H} contains the constant functions. And, if $f_n \in \mathcal{H}$ for $n \geq 1$ and $f_n \rightarrow f$ boundedly then, by the dominated convergence theorem, f is in \mathcal{H} as well. So \mathcal{H} is a linear subspace of bounded functions which includes the constant functions and is closed under bounded convergence. Consider now the π -class $\mathcal{P} = \mathcal{M} \otimes \mathcal{N}$. If $A \times B$ is in \mathcal{P} then $1_{A \times B}(x, y) = 1_A(x)1_B(y)$ and so the map $x \mapsto \int_{\mathcal{Y}} \nu(dy) 1_{A \times B}(x, y) = \nu(B)1_A(x)$ is \mathcal{M} -measurable, the map $y \mapsto \int_{\mathcal{X}} \mu(dx) 1_{A \times B}(x, y) = \mu(A)1_B(y)$ is \mathcal{N} -measurable, and

$$\int_{\mathcal{X}} \mu(dx) \int_{\mathcal{Y}} \nu(dy) 1_{A \times B}(x, y) = \mu(A)\nu(B) = \int_{\mathcal{Y}} \nu(dy) \int_{\mathcal{X}} \mu(dx) 1_{A \times B}(x, y).$$

Thus, $\chi(\mathcal{P}) \subseteq \mathcal{H}$. By the functional form of the π - λ theorem, it follows that $\chi(\mathcal{M} \otimes \mathcal{N}) \subseteq \mathcal{H}$ and so, if $C \in \mathcal{M} \otimes \mathcal{N}$ then $1_C \in \mathcal{H}$. By linearity, if $s = \sum_{j=1}^m s_j 1_{C_j}$ is simple with $C_j \in \mathcal{M} \otimes \mathcal{N}$ then $s \in \mathcal{H}$. Finally, if f is any positive, extended real-valued function, measurable with respect to $\mathcal{M} \otimes \mathcal{N}$, we may construct an increasing sequence of simple functions $\{s_n, n \geq 1\}$ with $s_n \uparrow f$ pointwise. Then $f \in \mathcal{H}$ as well by the monotone convergence theorem. This proves the theorem for the case of finite measures.

We extend the formulation to the case of σ -finite measures μ and ν by approximation by finite measures over larger and larger sets. Let $\{A_m, m \geq 1\}$ be an increasing sequence of sets in \mathcal{M} , each of finite μ -measure, and so that $A_m \uparrow \mathcal{X}$. For each m , these sets induce the finite measure $\mu_m(A) = \mu(A \cap A_m)$ or, in differential notation, $\mu_m(dx) = 1_{A_m}(x) \mu(dx)$. Likewise, let $\{B_n, n \geq 1\}$ be an increasing sequence of sets in

\mathcal{N} , each of finite ν -measure, and so that $B_n \uparrow \mathcal{Y}$. For each n , these sets induce the finite measure $\nu_n(B) = \nu(B \cap B_n)$, or, in differential notation, $\nu_n(dy) = 1_{B_n}(y)\nu(dy)$. Let f be any positive, extended real-valued function, measurable with respect to $\mathcal{M} \otimes \mathcal{N}$ and introduce the doubly indexed sequence of functions $f_{m,n}(x, y) = f(x, y)1_{A_m}(x)1_{B_n}(y)$. As we have proved that the theorem holds for finite measures, for each m and n , we see that the map $x \mapsto \int_y \nu(dy) f_{m,n}(x, y) = 1_{A_m}(x) \int_y \nu_n(dy) f(x, y)$ is \mathcal{M} -measurable, the map $y \mapsto \int_x \mu(dx) f_{m,n}(x, y) = 1_{B_n}(y) \int_x \mu_m(dx) f(x, y)$ is \mathcal{N} -measurable, and

$$\begin{aligned} \int_x \mu(dx) \int_y \nu(dy) f_{m,n}(x, y) &= \int_x \mu_m(dx) \int_y \nu_n(dy) f(x, y) \\ &= \int_y \nu_n(dy) \int_x \mu_m(dx) f(x, y) = \int_y \nu(dy) \int_x \mu(dx) f_{m,n}(x, y). \end{aligned}$$

Thus, the conclusion of the theorem holds for each $f_{m,n}$. As $f_{m,n}$ converges to f from below, we may allow $m \rightarrow \infty$ and $n \rightarrow \infty$ and by repeated applications of the monotone convergence theorem conclude that the theorem holds for f . \blacktriangleright

Say that a measure π on $\mathcal{M} \otimes \mathcal{N}$ is a *product measure* induced by the σ -finite measures μ and ν if

$$\pi(A \times B) = \mu(A)\nu(B) \quad (\text{for all } A \in \mathcal{M} \text{ and } B \in \mathcal{N}). \quad (5.1)$$

Theorem 1 now suggests a natural candidate for the product measure. Write $\mu \otimes \nu$ for the set function which to each $C \in \mathcal{M} \otimes \mathcal{N}$ assigns the value

$$\mu \otimes \nu(C) = \int_X \mu(dx) \int_Y \nu(dy) 1_C(x, y) = \int_Y \nu(dy) \int_X \mu(dx) 1_C(x, y). \quad (5.2)$$

COROLLARY *The set function $\mu \otimes \nu$ is the unique σ -finite product measure induced by the σ -finite measures μ and ν .*

PROOF: The proof is not at all hard if the reader has absorbed the $\pi\lambda$ theorem but may be deferred in the interests of getting to the punch line. The reader should begin with the simple verification that the set function $\mu \otimes \nu$ given by (5.2) indeed determines a σ -finite product measure on $\mathcal{M} \otimes \mathcal{N}$. The proof that this product measure is unique now follows the same pattern as that of Theorem XI.5.4. Suppose π_1 and π_2 are two product measures on $\mathcal{M} \otimes \mathcal{N}$ satisfying (5.1). Let $\mathcal{C} = \{C : \pi_1(C) = \pi_2(C)\}$ be the family of $\mathcal{M} \otimes \mathcal{N}$ -measurable sets on which π_1 and π_2 agree. By definition, the π -class $\mathcal{P} = \mathcal{M} \times \mathcal{N}$ is contained in \mathcal{C} and so it will suffice by the $\pi\lambda$ theorem to show that \mathcal{C} is a λ -class. It is clear that $X \times Y$ is in \mathcal{P} , hence in \mathcal{C} . Now suppose $\{C_n, n \geq 1\}$ is an increasing sequence of $\mathcal{M} \otimes \mathcal{N}$ -measurable sets and $C_n \uparrow C$. Then $\pi_1(C_n) \rightarrow \pi_1(C)$ and $\pi_2(C_n) \rightarrow \pi_2(C)$. But $\pi_1(C_n) = \pi_2(C_n)$ for each n and so the limits must coincide, $\pi_1(C) = \pi_2(C)$, and \mathcal{C} is closed under monotone unions. Finally, if C and D are elements of \mathcal{C} and $C \subseteq D$, then $\pi_1(D \setminus C) = \pi_1(D) - \pi_1(C) = \pi_2(D) - \pi_2(C) = \pi_2(D \setminus C)$ so that \mathcal{C} is closed under monotone set differences. Thus \mathcal{C} is a λ -class and hence $\mathcal{M} \otimes \mathcal{N} = \sigma(\mathcal{P}) \subseteq \mathcal{C}$. This forces π_1 and π_2 to coincide. \blacktriangleright

FUBINI'S THEOREM Suppose $(\mathcal{X}, \mathcal{M}, \mu)$ and $(\mathcal{Y}, \mathcal{N}, \nu)$ are σ -finite measure spaces, $(\mathcal{X} \times \mathcal{Y}, \mathcal{M} \otimes \mathcal{N}, \mu \otimes \nu)$ the corresponding product space equipped with product measure. If f is an integrable function on $\mathcal{X} \times \mathcal{Y}$ (with respect to product measure $\mu \otimes \nu$) then

$$\int_{\mathcal{X} \times \mathcal{Y}} f d(\mu \otimes \nu) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} f d\nu \right\} d\mu = \int_{\mathcal{Y}} \left\{ \int_{\mathcal{X}} f d\mu \right\} d\nu. \quad (5.3)$$

PROOF: The claimed result holds for indicator $f = 1_C$ by (5.2) and Theorem 1, hence by linearity for simple, positive f , hence by the monotone convergence theorem for positive f . Now suppose f is a generic integrable function with respect to product measure $\mu \otimes \nu$. Decompose f into its positive and negative parts, $f = f^+ - f^-$, where by definition, f^+ and f^- are both positive and integrable with respect to $\mu \otimes \nu$. By linearity,

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{Y}} f d(\mu \otimes \nu) &= \int_{\mathcal{X} \times \mathcal{Y}} f^+ d(\mu \otimes \nu) - \int_{\mathcal{X} \times \mathcal{Y}} f^- d(\mu \otimes \nu) \\ &= \int_{\mathcal{X}} d\mu \int_{\mathcal{Y}} d\nu f^+ - \int_{\mathcal{X}} d\mu \int_{\mathcal{Y}} d\nu f^- = \int_{\mathcal{X}} d\mu \int_{\mathcal{Y}} d\nu (f^+ - f^-) = \int_{\mathcal{X}} d\mu \int_{\mathcal{Y}} d\nu f. \end{aligned}$$

The second half of the claimed result is proved in exactly the same way. ▶

Fubini's theorem is just the natural analogue of the representation of a Riemann integral in the plane or in higher dimensions by an iterated integral. The reader should be wary, however, of reading more into Fubini's theorem than warranted. The counterexample that we began with illustrates the necessity of integrability with respect to product measure in Fubini's theorem: if f is integrable then, certainly, both of the integrals on the right in (5.3) exist, are equal, and coincide with the integral of f . But as we saw in our example, the existence of these iterated integrals is not in itself enough to guarantee that f is integrable with respect to $\mu \otimes \nu$; not even if the iterated integrals are equal—see Problem 9. It will suffice, however, if either iterated integral converges with f replaced by $|f|$.

THEOREM 3 *The existence of either of the iterated integrals*

$$\int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} |f| d\nu \right\} d\mu \quad \text{or} \quad \int_{\mathcal{Y}} \left\{ \int_{\mathcal{X}} |f| d\mu \right\} d\nu \quad (5.4)$$

implies that f is integrable (with respect to $\mu \otimes \nu$) and hence also the validity of (5.3).

The proof is a beautiful illustration of how a measure-theoretic viewpoint makes limiting arguments almost trite. The reader hungry for applications may push on, however, and return at leisure to absorb the argument.

PROOF: Suppose the first of the integrals on the left of (5.4) exists and is equal to M . We begin by dealing with domains in $\mathcal{X} \times \mathcal{Y}$ on which the measure of $\mu \times \nu$ is finite.

Suppose $\{\mathbb{A}_j, j \geq 1\}$ is an increasing sequence of $\mathcal{M} \otimes \mathcal{N}$ -measurable sets, each of finite $\mu \otimes \nu$ -measure, and such that $\mathbb{A}_j \uparrow \mathcal{X} \times \mathcal{Y}$. For each j and each $k \geq 1$, let $f_{j,k}(x, y) = \min\{k, |f(x, y)|1_{\mathbb{A}_j}(x, y)\}$. Then

$$\int_{\mathcal{X} \times \mathcal{Y}} f_{j,k} d(\mu \otimes \nu) \leq k\mu \otimes \nu(\mathbb{A}_j)$$

whence, for each j and k , $f_{j,k}$ is integrable and Fubini's theorem holds sway. It follows that

$$\int_{\mathcal{X} \times \mathcal{Y}} f_{j,k} d(\mu \otimes \nu) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} f_{j,k} d\nu \right\} d\mu \leq \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} |f| d\nu \right\} d\mu = M.$$

As, for each j , the sequence of functions $\{f_{j,k}, k \geq 1\}$ is increasing and converges pointwise to $|f|1_{\mathbb{A}_j}$, by the monotone convergence theorem, it follows that

$$\lim_{k \rightarrow \infty} \int_{\mathcal{X} \times \mathcal{Y}} f_{j,k} d(\mu \otimes \nu) = \int_{\mathcal{X} \times \mathcal{Y}} |f|1_{\mathbb{A}_j} d(\mu \otimes \nu) \leq M$$

and $|f|1_{\mathbb{A}_j}$ is integrable for each j . As the sequence of functions $\{|f|1_{\mathbb{A}_j}, j \geq 1\}$ increases pointwise to f it follows via another application of the monotone convergence theorem that

$$\lim_{j \rightarrow \infty} \int_{\mathcal{X} \times \mathcal{Y}} |f|1_{\mathbb{A}_j} d(\mu \otimes \nu) = \int_{\mathcal{X} \times \mathcal{Y}} |f| d(\mu \otimes \nu) \leq M,$$

and f is indeed integrable, whence (5.3) follows. The proof if the second member of (5.4) is finite follows along entirely similar lines. ▶

Specialising to probability measures, suppose $F(x)$ and $G(y)$ are d.f.s in Euclidean spaces \mathcal{X} and \mathcal{Y} , each in one or more dimensions, and let X and Y be independent random variables with these distributions. If $f(X, Y)$ is an integrable Baire function of the coordinate variables X and Y then

$$E(f(X, Y)) = \int_{\mathcal{X} \times \mathcal{Y}} f d(F \otimes G) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} f dG \right\} dF = \int_{\mathcal{Y}} \left\{ \int_{\mathcal{X}} f dF \right\} dG$$

and the expectation of f may be computed by iterated integrals.

Fubini's theorem sweeps measure-theoretic detail under the hood and a variety of useful consequences follow, almost effortlessly. Here, for instance, is a direct proof of Theorem 2.1 restated here for convenience.

THEOREM 4 *Independent integrable random variables are uncorrelated.*

PROOF: Suppose X and Y have marginal d.f.s F_1 and F_2 , respectively, engendering the product distribution $F = F_1 \otimes F_2$ of the pair (X, Y) . Then

$$\int_{\mathbb{R} \times \mathbb{R}} xy dF(x, y) = \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} xy dF_1(x) \right\} dF_2(y) = \left\{ \int_{\mathbb{R}} x dF_1(x) \right\} \left\{ \int_{\mathbb{R}} y dF_2(y) \right\},$$

or, compactly, $E(XY) = E(X) E(Y)$, by Fubini's theorem. ▶

The familiar *integration by parts* formula is another easy consequence.

THEOREM 5 Suppose g is a bounded, continuously differentiable Baire function on the line and suppose further that $\int_a^b |g'(x)| dx$ converges. Then

$$\int_{(a,b)} g(x) dF(x) = g(b)F(b-) - g(a)F(a) - \int_a^b g'(x)F(x) dx$$

for any open interval (a, b) (which may be infinite in one or both limits).

The reader should note that we need not be concerned with the integral limits on the right as the integral with respect to Lebesgue measure is insensitive to the addition or deletion of points on the boundary. The integral on the left is the expectation of $g1_{(a,b)}$ with respect to F . I will leave it to the reader to verify that if b is finite and the open interval (a, b) is replaced by the half-closed interval $(a, b]$ then a similar formula holds for the expectation of $g1_{(a,b]}$ with respect to F with the replacement of the term $F(b-)$ on the right by $F(b)$.

PROOF: Let \mathbb{A} be the collection of points (x, y) in the plane defined by the inequalities $a < x < b$ and $y \leq x$, and let $f(x, y) = |g'(x)| \cdot 1_{\mathbb{A}}(x, y)$. As $0 \leq F(x) \leq 1$,

$$\begin{aligned} \infty > \int_a^b |g'(x)| dx &\geq \int_a^b |g'(x)|F(x) dx \\ &= \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} f(x, y) dF(y) \right\} dx \end{aligned}$$

where we write $F(x) = \int_{(-\infty, x]} dF(y)$ to obtain the expression on the right. It follows that the iterated integral on the right converges whence, by Fubini,

$$\int_a^b g'(x)F(x) dx = \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} g'(x)1_{\mathbb{A}}(x, y) dF(y) \right\} dx = \int_{\mathbb{R}} \left\{ \int_{\mathbb{R}} g'(x)1_{\mathbb{A}}(x, y) dx \right\} dF(y).$$

As shown in Figure 3, the region of integration may be partitioned into two distinct regions: when $y \leq a$ the abscissa is constrained to lie in the range $a < x < b$, and when $a < y < b$ the abscissa can vary in the range $y < x < b$. Consequently,

$$\begin{aligned} \int_a^b g'(x)F(x) dx &= \int_{(-\infty, a]} \left\{ \int_a^b g'(x) dx \right\} dF(y) + \int_{(a,b)} \left\{ \int_y^b g'(x) dx \right\} dF(y) \\ &= \int_{(-\infty, a]} [g(b) - g(a)] dF(y) + \int_{(a,b)} [g(b) - g(y)] dF(y) \\ &= [g(b) - g(a)]F(a) + g(b)[F(b-) - F(a)] - \int_{(a,b)} g(y) dF(y). \end{aligned}$$

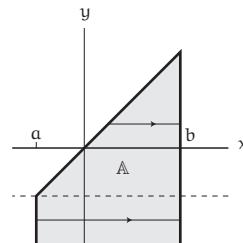


Figure 3: The region \mathbb{A} .

Collecting terms and reorganising the equation completes the proof. ▶

We will explore some of the ramifications of Fubini's theorem in the next few sections.

6 The volume of an n -dimensional ball

The blithe use of Fubini's theorem in converting integrals in the Euclidean plane to iterated one-dimensional integrals is so standard as not to require comment. The result can be even more devastating in impact in higher dimensions.

If $x = (x_1, \dots, x_n)$ is any point in \mathbb{R}^n we write $\|x\| = \sqrt{x_1^2 + \dots + x_n^2}$ for its Euclidean norm. Let $B^n(r) = \{x \in \mathbb{R}^n : \|x\| \leq r\}$ denote a ball of radius r in n dimensions and let $\text{Vol}(B^n(r))$ be its volume,

$$\text{Vol}(B^n(r)) = \int_{B^n(r)} dx = \int_{\substack{\dots \\ x_1^2 + \dots + x_n^2 \leq r^2}} dx_1 \dots dx_n.$$

Writing $V_n = \text{Vol}(B^n(1))$ for the volume of the n -dimensional unit ball, the change of variables $x_k \leftarrow x_k/r$ in the integral shows that $\text{Vol}(B^n(r)) = r^n V_n$.

For each n and r , let $\chi_{n,r}$ denote the indicator function for the ball $B^n(r)$. Then, for $n \geq 2$, we have

$$\chi_{n,1}(x_1, \dots, x_n) = \chi_{2,1}(x_1, x_2) \chi_{n-2, \sqrt{1-x_1^2-x_2^2}}(x_3, \dots, x_n),$$

whence an application of Fubini's theorem yields

$$V_n = \int_{\mathbb{R}^n} \chi_{n,1}(x) dx = \int_{B^2(1)} dx_1 dx_2 \int_{B^{n-2}(\sqrt{1-x_1^2-x_2^2})} dx_3 \dots dx_n.$$

We identify the inner integral with the volume of the ball $B^{n-2}(\sqrt{1-x_1^2-x_2^2})$ so that we obtain

$$\begin{aligned} V_n &= \int_{B^2(1)} \text{Vol}\left(B^{n-2}(\sqrt{1-x_1^2-x_2^2})\right) dx_1 dx_2 \\ &= V_{n-2} \int_{B^2(1)} (1-x_1^2-x_2^2)^{(n-2)/2} dx_1 dx_2. \end{aligned}$$

A change to polar coordinates, $\rho^2 = x_1^2 + x_2^2$, simplifies the integral over the unit disc $B^2(1)$ on the right and, via one more application of Fubini, results in

$$V_n = V_{n-2} \int_0^1 \int_0^{2\pi} (1-\rho^2)^{(n-2)/2} \rho d\theta d\rho \stackrel{(1-\rho^2=t^2)}{=} 2\pi V_{n-2} \int_0^1 t^{n-1} dt = \frac{2\pi V_{n-2}}{n}.$$

If we set $V_{-1} = 1/\pi$ and $V_0 = 1$ we can extend the recurrence to include the cases $n = 1$ and $n = 2$ as this gives $V_1 = 2\pi V_{-1} = 2$ and $V_2 = 2\pi V_0/2 = \pi$ in accordance with the length of the interval $[-1, 1]$ and the area of the unit circle, respectively. Churning through the recurrence then, when n is even we obtain

$$V_n = \frac{2^{n/2}\pi^{n/2}}{2 \cdot 4 \cdots (n-2)n} V_0 = \frac{\pi^{n/2}}{\frac{1}{2} \cdot \frac{3}{2} \cdots \left(\frac{n}{2}-1\right)\frac{n}{2}} \quad (n \text{ even}),$$

while when n is odd we obtain

$$V_n = \frac{2^{(n+1)/2}\pi^{(n+1)/2}}{1 \cdot 3 \cdots (n-2)n} V_{-1} = \frac{\pi^{(n-1)/2}}{\frac{1}{2} \cdot \frac{3}{2} \cdots \left(\frac{n-1}{2}-1\right)\frac{n-1}{2}} \quad (n \text{ odd}).$$

The results may be unified via the gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx \quad (t > 0). \quad (6.1)$$

If the reader does not already know it, she will be able to verify by an integration by parts that the gamma function satisfies the recurrence $\Gamma(t+1) = t\Gamma(t)$. As $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$ (an easy integration—see Problem VI.1), it is now easy to verify by induction that we have the following general solution.

THEOREM 1 *The volume of the unit ball in n dimensions is given by*

$$V_n = \int_{B^n(1)} d\mathbf{x} = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2} + 1\right)} \quad (n \geq 0).$$

As a quick check, we verify $V_0 = 1$ (a matter of definition), $V_1 = \sqrt{\pi}/\Gamma(3/2) = 2$, $V_2 = \pi/\Gamma(2) = \pi$, and $V_3 = \pi^{3/2}/\Gamma(5/2) = 4\pi/3$.

Our formulation now provides a way of tackling surface areas. The surface of the unit ball is denoted S^{n-1} and defined as the collection of points of unit length, that is to say the points $\mathbf{x} = (x_1, \dots, x_n)$ satisfying $\|\mathbf{x}\| = 1$. Now each non-zero point $\mathbf{x} = (x_1, \dots, x_n)$ may be uniquely represented in spherical coordinates by its Euclidean length $r = \|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}$ together with a directional parameter which we may take to be the unique point Ω at which the ray from the origin to the point \mathbf{x} intersects the surface S^{n-1} of the unit ball. It is not difficult to see that the Jacobian of the transformation $\mathbf{x} \mapsto (r, \Omega)$ is r^{n-1} (the reader desirous of dotting the i's can verify this as outlined in Problem 23) so that we may write a differential volume element as $d\mathbf{x} = r^{n-1} dr d\Omega$. Accordingly,

$$V_n = \int_{B^n(1)} d\mathbf{x} = \int_0^1 r^{n-1} dr \int_{S^{n-1}} d\Omega = \frac{1}{n} \int_{S^{n-1}} d\Omega,$$

and we have an accompanying result for surface areas.

THEOREM 2 *The surface area of the unit ball in n dimensions is given by*

$$A_n = \int_{S^{n-1}} d\Omega = \frac{n\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} \quad (n \geq 0).$$

An independent verification is always useful and we check to see that $A_2 = 2\pi$ and $A_3 = 4\pi$, as basic geometry tells us. (The reader may wish to ponder on what the relation $A_1 = 2$ says.)



7 The asymptotics of the gamma function

The gamma function plays an important rôle in analysis and crops up in unexpected places. Some insight into its behaviour may be obtained by combining expressions inside the integrand in (6.1) (replacing t by $t+1$ simplifies expressions) to obtain

$$\Gamma(t+1) = \int_0^\infty \exp(-x + t \log x) dx = t^{t+1} \int_0^\infty \exp(-t(u - \log u)) du,$$

the final step following via the natural change of variable $x = tu$. Differentiation shows that the function $u - \log u$ has a unique minimum of 1 attained at $u = 1$. A final change of variable $u = v + 1$ centres the integrand at the origin and results in the expression

$$\Gamma(t+1) = t^{t+1} e^{-t} \int_{-1}^\infty \exp\{-t(v - \log(v+1))\} dv.$$

We are hence led to consider the behaviour of the integral

$$I(t) = \int_{-1}^\infty e^{-tf(v)} dv$$

where f is the continuous function defined for $v > -1$ by $f(v) = v - \log(v+1)$. Easy differentiations show that $f'(v) = 1 - 1/(v+1)$ and $f''(v) = 1/(v+1)^2$ and it follows quickly that f increases monotonically on either side away from its single minimum of 0 at the origin. In view of the rapid extinction of the exponential in the integrand, for large values of t we hence anticipate that most of the contribution to the integral $I(t)$ will accrue from a small neighbourhood of the origin. This is the principle behind *Laplace's method of integration*. The reader who has ploughed through Section VI.6 will have seen the idea in action in a related setting. Now to put this observation to work.

For any $0 < \delta < 1$, there exists $\eta = \eta(\delta) > 0$ [which we may take to be the smaller of the values $f(-\delta)$ and $f(\delta)$] such that $f(v) > \eta$ for all $|v| \geq \delta$. Consequently,

$$tf(v) = (t-1)f(v) + f(v) > (t-1)\eta + f(v)$$

whenever $|v| \geq \delta$. It follows that

$$\begin{aligned} \int_{-1 < v \leq -\delta \text{ or } v \geq \delta} e^{-tf(v)} dv &< e^{-(t-1)\eta} \int_{-1 < v \leq -\delta \text{ or } v \geq \delta} e^{-f(v)} dv \\ &< e^{-(t-1)\eta} \int_{-1}^\infty e^{-f(v)} dv = e^{-(t-1)\eta} \int_{-1}^\infty (v+1)e^{-v} dv = e^{-t\eta} \cdot e^{\eta+1} \end{aligned} \quad (7.1)$$

and the contribution to the integral $I(t)$ away from the origin is exponentially subdominant for large t .

Now for the contribution near the origin. As $f(0) = f'(0) = 0$ and $f''(0) = 1$ we expect a Taylor expansion of f truncated at the quadratic term to provide an accurate approximation of f in a sufficiently small neighbourhood of the origin. To verify this we consider the function

$$g(v) = f(v) - f(0) - f'(0)v - f''(0)v^2/2 = f(v) - v^2/2.$$

It is easy to see that $g(0) = g'(0) = g''(0) = 0$ and, in particular, the fact that $g''(0) = 0$ implies that

$$\frac{g'(v) - g'(0)}{v} = \frac{g'(v)}{v} \rightarrow 0 \quad (\text{as } v \rightarrow 0).$$

It follows that, for any $\epsilon > 0$, we may select $\delta = \delta(\epsilon) > 0$ sufficiently small so that $|g'(v)| < \epsilon|v|$ for all $|v| < \delta$. In such a neighbourhood of the origin the mean value theorem tells us that $g(v) = g(0) + g'(v_0)v = g'(v_0)v$ for some v_0 lying between 0 and v . We consequently have $|g(v)| = |g'(v_0)| \cdot |v| < \epsilon v^2$ for all $|v| < \delta$. It follows hence that $\frac{1}{2}(1-2\epsilon)v^2 < f(v) < \frac{1}{2}(1+2\epsilon)v^2$ for all $|v| < \delta$ and we may bound the contribution to the integral $I(t)$ from the vicinity of the origin by

$$\int_{-\delta}^{\delta} e^{-(1+2\epsilon)t v^2/2} dv < \int_{-\delta}^{\delta} e^{-t f(v)} dv < \int_{-\delta}^{\delta} e^{-(1-2\epsilon)t v^2/2} dv.$$

Selecting $0 < \epsilon < 1/2$ to keep the exponents negative, we may expand the range of integration for the bookend integral bounds to go from $-\infty$ to $+\infty$ and, in view of the normal tail bound of Lemma VI.1.3, incur only a penalty

$$\int_{|v| \geq \delta} e^{-(1 \pm 2\epsilon)t v^2/2} dv < \frac{\sqrt{2\pi}}{\sqrt{t(1 \pm 2\epsilon)}} \cdot e^{-(1 \pm 2\epsilon)t \delta^2/2} \quad (7.2)$$

which is also exponentially subdominant for large t . (An intermediate step helps standardise the integral by changing variable to $w = v\sqrt{(1 \pm 2\epsilon)t}$.) We are finally left with a standard Gaussian integral corresponding to a zero-mean normal variable with variance $1/(1 \pm 2\epsilon)t$ and, via the same change of variable, we obtain

$$\int_{-\infty}^{\infty} e^{-(1 \pm 2\epsilon)t v^2/2} dv = \frac{\sqrt{2\pi}}{\sqrt{t(1 \pm 2\epsilon)}} \int_{-\infty}^{\infty} \phi(w) dw = \sqrt{\frac{2\pi}{t}} (1 \pm 2\epsilon)^{-1/2}.$$

The bounds on the right of (7.1) and (7.2) remain exponentially subdominant when divided by $\sqrt{2\pi/t}$ and, for sufficiently large t , each of these normalised bounds becomes less than ϵ . As, for all sufficiently small ϵ , and certainly for all ϵ in the interval $(0, 0.309)$,

$$1 - 2\epsilon \leq (1 + 2\epsilon)^{-1/2} \leq (1 - 2\epsilon)^{-1/2} \leq 1 + 2\epsilon,$$

we may divide throughout by $\sqrt{2\pi/t}$ and, collecting the miscellaneous subdominant terms that we have dropped by the wayside, obtain

$$1 - 4\epsilon < \frac{I(t)}{\sqrt{2\pi/t}} < 1 + 4\epsilon$$

for all sufficiently large t . As ϵ may be taken arbitrarily small, it follows that $I(t)/\sqrt{2\pi/t} \rightarrow 1$ as $t \rightarrow \infty$.

THEOREM As $t \rightarrow \infty$,

$$\frac{\Gamma(t+1)}{\sqrt{2\pi} t^{t+1/2} e^{-t}} \rightarrow 1.$$

By allowing t to approach infinity through integer values n , we obtain

$$n! \sim \sqrt{2\pi} n^{n+1/2} e^{-n} \quad (n \rightarrow \infty). \quad (7.3)$$

This is Stirling's celebrated formula for the factorial.

EXAMPLES: 1) The central term of the binomial provides a quick illustration of the utility of the result. As

$$b_{2n}(n; 1/2) = 2^{-2n} \binom{2n}{n} = \frac{2^{-2n} (2n)!}{(n!)^2},$$

two applications of Stirling's formula to the terms $(2n)!$ and $n!$ (and some careful book-keeping) shows that

$$\frac{b_{2n}(n; 1/2)}{\sqrt{1/(n\pi)}} \rightarrow 1 \quad (n \rightarrow \infty), \quad (7.4)$$

as the reader may recall seeing derived (with $2n$ replaced by n) in VI.6.7 by another Laplace argument via the correspondence of binary digits and Rademacher functions. Thus, in 100 tosses of a fair coin, the probability of the most probable outcome—that heads and tails are even—is approximately $\sqrt{1/(50\pi)} = 0.079788 \dots$. The quality of the approximation is made clear by comparison with the exact value $b_{100}(50; 1/2) = 0.079589 \dots$.

2) As another application, the volume of the unit ball in a large number of dimensions has a growth rate

$$V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)} \sim \frac{1}{\sqrt{\pi n}} \left(\frac{2\pi e}{n} \right)^{n/2}.$$

It follows that $V_n \rightarrow 0$ superexponentially fast. (The largest volume is at $n = 5$ with $V_5 = 5.26379 \dots$. It's all downhill from there.) ▶

8 A question from antiquity

In the mythical story of the founding of Carthage, Dido, a princess of Tyre, flees from her tyrannical brother, Pygmalion, and, with the help of a sympathetic crew, takes to the high seas. After many adventures she lands, bedraggled and forlorn, on the north coast of Africa. The natives are none too happy to see Dido and her retinue, but when Dido proposes to them that they cede her a small portion of land—only as much as can be contained by an ox hide—for which she will pay handsomely, they can't refuse the deal. More fool this royal simpleton, think they. Dido, as sharp as she is wilful, has a large ox slaughtered and its hide cut into thin strips which, knotted together, form a long rope. The land circumscribed by the ox hide rope becomes the foundation of the ancient kingdom of Carthage, Dido its legendary queen. Like many, I find the story delightful and cannot resist providing a facsimile of Mathäus Merian, the Elder's beautiful visualisation of the scene in Figure 4.



Figure 4: Engraving of Dido of Tyre by Mathäus Merian, the Elder, 1630. This image is credited by Rodrigo Bañuelos to Hilderbrandt-Troma, “Shape and Form of the Natural Universe”, though I have been unable to trace the reference.

The charming story of Dido centres on a key geometrical question: what shape should the land circumscribed by the rope take? The shape should clearly be chosen to maximise the amount of land that it contains.

THE ISOPERIMETRIC PROBLEM *Of all closed curves of a given length L, which one encloses the largest area A?*

It is apparent that a dual formulation of this problem can be posed.

THE AUTHALIC PROBLEM *Of all closed curves enclosing a given area A, which one has the smallest perimeter L?*

The answer to both problems is inevitably the circle though the demonstration is far from trivial. The principle generalises naturally enough to any number of dimensions, balls replacing circles in the resolution.

The isoperimetric principle has been known since antiquity—and continues to enthral. Dido's excellent adventure, embellished and somewhat grown

in the telling, perhaps, is traditionally dated to around 800 B.C., although archaeological evidence gleaned from Carthaginian ruins may make the date suspect. In ancient Rome the quaestor Proclus outlined rules for inheritance in socialist communes so that each family received a plot of the same perimeter. In agriculture, a farmer with a given perimeter wishes to maximise the arable area. Faced with the problem of guarding a fixed population (area) military encampments would wish to expose the smallest perimeter to attack. Settlers in a Martian colony would be faced with the design of living quarters which maximise space (volume) given scarce resources (perimeter).

The reader who looks at an old map will find that ancient Carthage was established not in the interior but on the coast of North Africa in what is now modern-day Tunisia. Indeed, Virgil reports in the *Aeneid* (written about 29 B.C.) that Dido increased the land under her control by using the coastline as a fixed boundary and annexing the land between her ox hide rope and the sea. Here then is a modified isoperimetric problem: *determine the figure with the largest area bounded on one side by a straight line (the coastline) and on the other by a curve of given length (the ox hide)*. Virgil reports that Dido acquired between 8 and 32 hectares of land by this stratagem. The reader may find some diversion and amusement in checking Virgil's calculations by positing an unfortunate cylindrical ox of girth 3 metres and height 2 metres whose hide is cut into strips of width 0.5 centimetres.

The elementary isoperimetric theorem has had, and continues to have, an abiding impact on an incredible number of areas in mathematics, science, and engineering. In probability, investigations into Gaussian isoperimetry were instigated by Paul Lévy and led to very deep consequences. I will consider a simple related problem here and return to the general related framework of concentration inequalities in Chapter XVII.

Consider \mathbb{R}^n equipped with the normal product measure $\Phi^{\otimes n}$. A random vector $Z = (Z_1, \dots, Z_n)$ drawn according to $\Phi^{\otimes n}$ then has the normal density

$$f(z) = \prod_{k=1}^n \phi(z_k) = (2\pi)^{-n/2} e^{-\|z\|^2/2}.$$

The simple observation that the density f decays monotonically radially away from the origin has important consequences. For any $V \geq 0$ we adopt the nonce notation

$$r(V) = \pi^{-1/2} (\Gamma(n/2 + 1)V)^{1/n}$$

for the radius of a ball of volume V and write $B = B^n(r(V))$ in short for the ball of volume V centred at the origin.

AN ISOPERIMETRIC THEOREM *With P identified with product Gaussian measure $\Phi^{\otimes n}$, among all Borel sets of a given volume, the ball centred at the origin has the*

largest probability. In notation, $\sup \mathbf{P}\{\mathbf{Z} \in \mathbb{A}\} = \mathbf{P}\{\mathbf{Z} \in \mathbb{B}\}$ where the supremum is over all Borel sets \mathbb{A} of volume V .

Indeed, if \mathbb{A} has volume V then the sets $\mathbb{A} \setminus \mathbb{B}$ and $\mathbb{B} \setminus \mathbb{A}$ have the same volume or, in notation, $\int_{\mathbb{A} \setminus \mathbb{B}} dz = \int_{\mathbb{B} \setminus \mathbb{A}} dz$. Any differential volume element dz in $\mathbb{A} \setminus \mathbb{B}$ is at radial distance $> r(V)$ from the origin while any corresponding differential element dz in $\mathbb{B} \setminus \mathbb{A}$ is at radial distance $\leq r(V)$ from the origin; see Figure 5. As

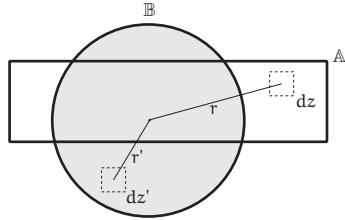


Figure 5: Gaussian isoperimetry.

$f(z)$ decreases monotonically with the radial distance $\|z\|$, it follows that

$$\int_{\mathbb{A} \setminus \mathbb{B}} f(z) dz \leq (2\pi)^{-n/2} e^{-r(V)^2/2} \text{Vol } \mathbb{A} \setminus \mathbb{B} = (2\pi)^{-n/2} e^{-r(V)^2/2} \text{Vol } \mathbb{B} \setminus \mathbb{A} \leq \int_{\mathbb{B} \setminus \mathbb{A}} f(z) dz,$$

and the claimed result follows.

More generally, for $1 \leq j \leq M$, suppose \mathbb{A}_j is a Borel set of volume V_j and let $V = \frac{1}{M}(V_1 + \dots + V_M)$ be the arithmetic mean of their volumes. In view of the isoperimetric theorem, probabilities increase if we replace each set \mathbb{A}_j by a ball \mathbb{B}_j of volume V_j centred at the origin, and it follows that

$$\frac{1}{M} \sum_{j=1}^M \mathbf{P}\{\mathbf{Z} \in \mathbb{A}_j\} \leq \frac{1}{M} \sum_{j=1}^M \mathbf{P}\{\mathbf{Z} \in \mathbb{B}_j\}. \quad (8.1)$$

If each of the sets \mathbb{A}_j has volume V then, for each j , $\mathbb{B}_j = \mathbb{B}$ is the ball of volume V centred at the origin and the bound on the right becomes $\mathbf{P}\{\mathbf{Z} \in \mathbb{B}\}$. If not all the balls \mathbb{B}_j have the same volume then, as shown in Figure 6, we may transfer

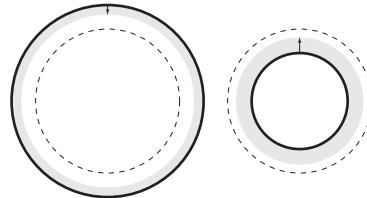


Figure 6: Symmetrisation increases Gaussian probabilities.

volume from a larger ball to a smaller ball creating two balls more nearly of a size with each other and, without altering the total volume of the balls, increase the arithmetic mean of their probabilities. By a repeated systematic process of symmetrisations of this type we may then reduce the system of balls to the case where all balls have the same volume V . In detail, suppose there exist balls \mathbb{B}_i and \mathbb{B}_k with volumes $V_i < V < V_k$. Let $V_0 = \min\{V - V_i, V_k - V\}$. We now replace the system of balls $\{\mathbb{B}_j\}$ by a new system of balls $\{\mathbb{B}'_j\}$ by replacing \mathbb{B}_i and \mathbb{B}_k by balls \mathbb{B}'_i and \mathbb{B}'_k of volumes $V'_i = V_i + V_0$ and $V'_k = V_k - V_0$, respectively, and keeping the remaining balls intact, $\mathbb{B}'_j = \mathbb{B}_j$ if j is not equal to i or k . It is clear that the arithmetic mean of the volumes of the new system of balls $\{\mathbb{B}'_j\}$ remains V . As

$$\int_{\mathbb{B}'_i \setminus \mathbb{B}_i} f(z) dz \geq (2\pi)^{-n/2} e^{-r(V)^2/2} V_0 \geq \int_{\mathbb{B}_k \setminus \mathbb{B}'_k} f(z) dz$$

by the radially monotonic nature of the normal density f , by comparing the arithmetic means of the probabilities of the two systems of balls, we find

$$\frac{1}{M} \sum_{j=1}^M \mathbf{P}\{\mathbf{Z} \in \mathbb{B}'_j\} - \frac{1}{M} \sum_{j=1}^M \mathbf{P}\{\mathbf{Z} \in \mathbb{B}_j\} = \frac{1}{M} \left[\int_{\mathbb{B}'_i \setminus \mathbb{B}_i} f(z) dz - \int_{\mathbb{B}_k \setminus \mathbb{B}'_k} f(z) dz \right] \geq 0.$$

Thus, we have now created a new system of balls without changing the total volume, the arithmetic mean of whose probabilities is larger. This new system contains at least one new ball whose volume is exactly equal to the arithmetic mean V of the volumes. By repeating this procedure we may systematically increase the number of balls of volume V without changing the total volume and while increasing the arithmetic mean of the probabilities. This sequence of symmetrisations will terminate in no more than $M - 1$ steps with all balls having volume V .

ISOPERIMETRIC THEOREM, REFINEMENT *With \mathbf{P} identified with product Gaussian measure $\Phi^{\otimes n}$, the arithmetic mean of the probabilities of any finite system of Borel sets of finite volume is maximised when each of the Borel sets is the ball centred at the origin with volume equal to the arithmetic mean of the volumes of the sets. In notation, $\sup \frac{1}{M} \mathbf{P}\{\mathbf{Z} \in \mathbb{A}_j\} = \mathbf{P}\{\mathbf{Z} \in \mathbb{B}\}$ where the supremum is over all Borel sets $\mathbb{A}_1, \dots, \mathbb{A}_M$ the arithmetic mean of whose volumes is V .*

The reader will be able to see echoes of the isoperimetric problem in our Gaussian isoperimetry principle. In spite of its trivial-seeming nature it leads to deep and unexpected results as the reader will see in the following section.

9 How fast can we communicate?

We are grown inured to the wonders of modern wireless communications. Data are beamed to satellites from cell phone towers and back again, seamlessly,

transparently, and at wondrous speeds. But as recently as 1948 it was not apparent what could be done.

In an only slightly sanitised setting we suppose that a real value can be transmitted in each use of a communication medium. In n uses of the medium then one can transmit a real *signal* vector $\mathbf{s} = (s_1, \dots, s_n)$. Two fundamental constraints limit the precision with which the receiver can recover the transmitted signal vector. On the one hand, imperfections in the medium result in the garbling of transmissions by the addition of an independent Gaussian *noise* term to each transmission. With $\mathbf{Z} = (Z_1, \dots, Z_n)$ a sequence of independent variables drawn from the common normal distribution with mean zero and variance P_N , the received vector after n transmissions, $\mathbf{X} = \mathbf{s} + \mathbf{Z}$, is hence an additively corrupted version of the transmitted signal vector \mathbf{s} . On the other hand, resource constraints limiting the available transmission power force the selected signal vector to satisfy the inequality $\frac{1}{n}\|\mathbf{s}\|^2 \leq P_S$. The quantities P_S and P_N represent the mean power of the signal and noise processes and play a fundamental rôle. This is the setting of a *power-limited Gaussian communication channel*, satellite channels providing the archetypal example.

Given a finite set of M possible messages a_1, \dots, a_M to be communicated, the systems designer associates with each message a_i a unique signal vector $\mathbf{s}_i \in \mathbb{R}^n$ satisfying the mean signal power constraint. The constellation of signal vectors $\{\mathbf{s}_1, \dots, \mathbf{s}_M\}$ forms a *code book* for transmission, each signal vector acting as a surrogate for the associated message. Both sender and receiver are made aware of the selected code book before communication is begun.

There is no great loss of generality in this framework. The reader, accustomed as she is to the modern digital world, will have no difficulty in accepting that by digitisation we may satisfactorily represent the distinguishable entities that constitute possible messages in a given application domain by a, possibly very large, finite set of possibilities. Assuming that a certain latency is permissible even a large set of message possibilities may be accommodated by very many uses of the medium, that is to say, by a constellation of signal vectors in a large number n of dimensions.

Suppose now that one of the M messages is selected for transmission. The sender through n uses of the channel transmits the associated signal vector; the receiver, of course, does not know *a priori* which, else there would be little point in the transmission. At an intuitive level, if the signal vectors are close to each other then we feel that perturbation caused by random noise will make it impossible for the receiver to reliably distinguish between transmissions of different signals. This may suggest to us that we try to separate the signal vectors as much as possible. The signal power constraint, however, limits how far apart we can place them. While in one dimension it is clear that, if the signals are equally likely to be selected for transmission, we should equispace the signals between the locations $-\sqrt{P_S}$ and $+\sqrt{P_S}$, it is not at all clear how to proceed in many dimensions, nor what performance we can expect.

How many differentiated messages then can be reliably communicated by n uses of the medium? Our criterion for reliability is quite stringent: we require that, for any $\epsilon > 0$, the signalling strategy permits the recovery of any of the M transmitted signal vectors with probability exceeding $1 - \epsilon$. As each of the signal vectors may be identified, ignoring integer round-off, by a sequence of $\log_2 M$ binary digits or *bits* [a bit, by convention, is an element of $\{0, 1\}$], say by a lexicographic ordering, a transmission of a message through n uses of the channel may be thought of conceptually as transmitting at a rate $R = \frac{1}{n} \log_2 M$ bits per dimension. For a fixed rate R , an ever larger number of messages $M = 2^{nR}$ can be accommodated by increasing the dimensionality. Our basic question then becomes: *what is the largest rate (in bits per dimension) at which information can be reliably communicated?*

As noise seems to place an ineluctable limitation on the precision of communication we should begin by understanding its impact.

As the noise components are independent, the vector $Z = (Z_1, \dots, Z_n)$ has the spherically symmetric normal density $f(z) = P_N^{-n/2} e^{-\|z\|^2/2P_N}$. It is natural now to consider the Euclidean length of Z . Now, with vectors identified as *row vectors*, $\|Z\|^2 = ZZ^T = Z_1^2 + \dots + Z_n^2$. [The reader will recognise from Section X.2 that $\|Z\|^2$ has the chi-squared density $P_N^{-1} g_{n/2}(P_N^{-1} t; 1/2)$ but it will be more convenient for our purposes to deal directly with the originating density $f(z)$.] As the noise components arise from the same normal distribution $\mathcal{N}(0, P_N)$, additivity of expectation shows that $E(\|Z\|^2) = n E(Z_1^2) = n \text{Var}(Z_1) = n P_N$, and, as the noise components are independent, we also have

$$\text{Var}(\|Z\|^2) = n \text{Var}(Z_1^2) = n(E(Z_1^4) - E(Z_1^2)^2) = n(3P_N^2 - P_N^2) = 2nP_N^2$$

in view of Lemma VI.1.2. As the standard deviation of $\|Z\|^2$ increases only as the square-root of the number of dimensions we may anticipate that the squared length of the noise vector gets increasingly concentrated around its mean value. We are accordingly led to consider the probability of a proportionally small deviation from the mean

$$\begin{aligned} P\{|\|Z\|^2 - nP_N| > \epsilon n\} \\ &= \int_{|\|z\|^2 - nP_N| > \epsilon n} f(z) dz \leq \int_{|\|z\|^2 - nP_N| > \epsilon n} \left(\frac{\|z\|^2 - nP_N}{\epsilon n} \right)^2 f(z) dz \\ &\leq \int_{\mathbb{R}^n} \left(\frac{\|z\|^2 - nP_N}{\epsilon n} \right)^2 f(z) dz = \frac{\text{Var}(\|Z\|^2)}{\epsilon^2 n^2} = \frac{2P_N^2}{\epsilon^2 n} < \epsilon \quad (\text{eventually}). \end{aligned} \tag{9.1}$$

The reader may feel on reflection that the bounding process appears familiar. With cause—this is the inequality of Chebyshev that we used to bound binomial probabilities in Section V.6 in the proof of the law of large numbers.

Thus, for any $\epsilon > 0$, with probability at least $1 - \epsilon$ we have

$$n(P_N - \epsilon) \leq \|Z\|^2 \leq n(P_N + \epsilon),$$

and $\|Z\|$ is confined with high probability to a thin shell at the boundary of the ball of radius $\sqrt{n P_N}$. This phenomenon is sometimes referred to as *sphere-hardening*: the soap film at the boundary of the ball of radius $\sqrt{n P_N}$ contains most of the probability in high dimensions.

Now suppose that a signal vector $s \in \mathbb{B}^n(\sqrt{n P_S})$ is transmitted. The received vector $X = s + Z$ has density $g(x) = f(x - s)$ so that the previous argument shows that X is concentrated near the boundary of the ball of radius $\sqrt{n P_N}$ centred at s . As s satisfies the mean power constraint $\|s\|^2 \leq n P_S$, we anticipate hence that X will be confined within a ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$. Verification follows a very similar pattern. As

$$\|X\|^2 = (s + Z)(s + Z)^T = \|s\|^2 + 2sZ^T + \|Z\|^2,$$

additivity of expectation shows that

$$\mathbb{E}(\|X\|^2) = \|s\|^2 + 2s \mathbb{E}(Z^T) + \mathbb{E}(\|Z\|^2) = \|s\|^2 + n P_N.$$

In view of the elementary inequality $(a + b)^2 \leq 2a^2 + 2b^2$, we also have

$$\begin{aligned} \text{Var}(\|X\|^2) &= \mathbb{E}((\|X\|^2 - \|s\|^2 - n P_N)^2) = \mathbb{E}((2sZ^T + \|Z\|^2 - n P_N)^2) \\ &\leq \mathbb{E}(2(2sZ^T)^2 + 2(\|Z\|^2 - n P_N)^2) = 8s \mathbb{E}(Z^T Z)s^T + 2 \text{Var}(\|Z\|^2). \end{aligned}$$

As $\mathbb{E}(Z^T Z) = P_n I_n$ where I_n is the identity matrix of order n and $\text{Var}(\|Z\|^2) = 2n P_n^2$, we obtain the bound

$$\text{Var}(\|X\|^2) \leq 8\|s\|^2 P_N + 4n P_n^2 \leq 8n P_S P_N + 4n P_n^2 = 4n P_N(2P_S + P_N),$$

and again the variance increases only linearly with dimension n . By another deployment of the Chebyshev argument we see then that

$$\begin{aligned} \mathbb{P}\{\|X\|^2 > n(P_S + P_N + \epsilon)\} &= \mathbb{P}\{\|X\|^2 - \|s\|^2 - n P_N > n P_S - \|s\|^2 + n \epsilon\} \\ &\leq \mathbb{P}\{\|X\|^2 - \|s\|^2 - n P_N > n \epsilon\} = \int_{\|\mathbf{x}\|^2 - \|\mathbf{s}\|^2 - n P_N > n \epsilon} f(\mathbf{x} - \mathbf{s}) d\mathbf{x} \\ &\leq \int_{\|\mathbf{x}\|^2 - \|\mathbf{s}\|^2 - n P_N > n \epsilon} \left(\frac{\|\mathbf{x}\|^2 - \|\mathbf{s}\|^2 - n P_N}{n \epsilon} \right)^2 f(\mathbf{x} - \mathbf{s}) d\mathbf{x} \\ &\leq \int_{\mathbb{R}^n} \left(\frac{\|\mathbf{x}\|^2 - \|\mathbf{s}\|^2 - n P_N}{n \epsilon} \right)^2 f(\mathbf{x} - \mathbf{s}) d\mathbf{x} = \frac{\text{Var}(\|X\|^2)}{n^2 \epsilon^2} \leq \frac{4 P_N(2P_S + P_N)}{n \epsilon^2} < \epsilon, \quad (9.2) \end{aligned}$$

eventually for sufficiently large n . It follows that for all choices of signal vectors satisfying the mean power constraint, with high probability at least $1 - \epsilon$, the received vector X is confined asymptotically to the ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$ centred at the origin.

Thus, as a consequence of sphere-hardening, when a signal vector \mathbf{s} is transmitted, the received vector \mathbf{X} migrates at least a distance $\sqrt{n(P_N - \epsilon)}$ from \mathbf{s} , but stays confined within the ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$ centred at the origin. This suggests that in order to be reliably identified, signal vectors must be at least a distance $\sqrt{n(P_N - \epsilon)}$ from each other, or, in other words, the balls of radius $\sqrt{n(P_N - \epsilon)}$ centred at each of the M signal vectors in the code book do not intersect. But geometry limits how many small balls can be packed into a larger ball. Indeed, the number of balls of radius $\sqrt{n(P_N - \epsilon)}$ that can be packed into a ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$ is certainly less than

$$\frac{\text{Vol } \mathbb{B}^n(\sqrt{n(P_S + P_N + \epsilon)})}{\text{Vol } \mathbb{B}^n(\sqrt{n(P_N - \epsilon)})} = \frac{(n(P_S + P_N + \epsilon))^{n/2}}{(n(P_N - \epsilon))^{n/2}} = \left(1 + \frac{P_S + 2\epsilon}{P_N - \epsilon}\right)^{n/2} \quad (9.3)$$

and the expression on the right then appears to be an upper bound on how many distinct messages can be communicated by n uses of the medium. A fully formal argument does little more than colour in this geometric picture.

The receiver must make a decision on what was transmitted based entirely upon the received vector and the preselected code book of possibilities for the signal vector. The operation of the receiver is captured by a *decision function* $\hat{\mathbf{s}}: \mathbb{R}^n \rightarrow \{\mathbf{s}_1, \dots, \mathbf{s}_M\}$ which maps each realisation \mathbf{x} of the received vector to a signal vector $\hat{\mathbf{s}}(\mathbf{x})$ in the code book. The decision function $\hat{\mathbf{s}}$ enjoins a partition of \mathbb{R}^n into M disjoint regions $\mathbb{I}_1, \dots, \mathbb{I}_M$, with $\hat{\mathbf{s}}: \mathbf{x} \mapsto \mathbf{s}_j$ if, and only if, $\mathbf{x} \in \mathbb{I}_j$. It will suffice for our purposes to assume that the regions \mathbb{I}_j form Borel sets though it is not difficult to see that the decision regions minimising the mean probability of receiver error form convex polyhedra whose boundaries are the hyperplanes in \mathbb{R}^n orthogonal to the lines connecting each pair of signal vectors; see Figure 7.

Suppose now that a signal vector \mathbf{s}_j from a given code book $\{\mathbf{s}_1, \dots, \mathbf{s}_M\}$ has been transmitted. The receiver sees the additively corrupted version $\mathbf{X} = \mathbf{s}_j + \mathbf{Z}$ and produces $\hat{\mathbf{s}}(\mathbf{X})$ as her estimate of what was transmitted, her estimate being correct if, and only if, the received vector \mathbf{X} lies in the decision region \mathbb{I}_j associated with the signal vector \mathbf{s}_j . Let \mathbb{I}'_j be that portion of the region \mathbb{I}_j that lies within the ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$ centred at the origin and let $\mathbb{I}''_j = \mathbb{I}_j \setminus \mathbb{I}'_j$ be that portion of \mathbb{I}_j that lies outside this ball. Let V_j be the volume of \mathbb{I}'_j and let $r(V_j)$ be the radius of the ball \mathbb{B}_j of volume V_j centred at the origin. Then \mathbf{X} lies in \mathbb{I}'_j if, and only if, \mathbf{Z} lies in $\mathbb{I}'_j - \mathbf{s}_j$, the affine shift of \mathbb{I}'_j consisting of those points of the form $\mathbf{x} - \mathbf{s}_j$ with $\mathbf{x} \in \mathbb{I}'_j$. As volumes are unaffected by translations, the volume of the set $\mathbb{I}'_j - \mathbf{s}_j$ is V_j and, in view of the isoperimetric principle of the previous section, we then have $P\{\mathbf{X} \in \mathbb{I}'_j\} = P\{\mathbf{Z} \in \mathbb{I}'_j - \mathbf{s}_j\} \leq P\{\mathbf{Z} \in \mathbb{B}_j\}$. On

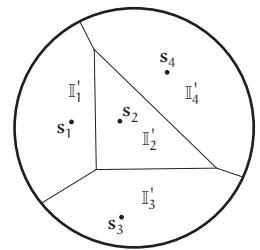


Figure 7: Power-limited decision regions for a code book.

the other hand, $\mathbf{P}\{\mathbf{X} \in \mathbb{I}_j''\} \leq \mathbf{P}\{\|\mathbf{X}\|^2 > n(P_S + P_N + \epsilon)\} < \epsilon$, eventually, by (9.2). It follows that the probability of the event $\widehat{s}(\mathbf{X}) = s_j$ that the receiver makes the correct decision may be bounded by

$$P_j = \mathbf{P}\{\mathbf{X} \in \mathbb{I}_j\} = \mathbf{P}\{\mathbf{X} \in \mathbb{I}_j'\} + \mathbf{P}\{\mathbf{X} \in \mathbb{I}_j''\} \leq \mathbf{P}\{\mathbf{Z} \in \mathbb{B}_j\} + \epsilon$$

for all sufficiently large n .

Let V denote the arithmetic mean of the volumes V_1, \dots, V_M and, as in the previous section, let $\mathbb{B} = \mathbb{B}^n(r(V))$ be the ball of volume V centred at the origin, $r(V)$ its radius. As the sets $\mathbb{I}_1', \dots, \mathbb{I}_M'$ partition the ball of radius $\sqrt{n(P_S + P_N + \epsilon)}$, the Gaussian isoperimetric principle tells us that the arithmetic mean of the probabilities P_j is bounded by

$$\frac{1}{M} \sum_{j=1}^M P_j \leq \frac{1}{M} \sum_{j=1}^m \mathbf{P}\{\mathbf{Z} \in \mathbb{B}_j\} + \epsilon \leq \mathbf{P}\{\mathbf{Z} \in \mathbb{B}\} + \epsilon.$$

If M exceeds the right-hand side of (9.3) then $r(V) \leq \sqrt{n(P_N - \epsilon)}$ and, by the sphere-hardening inequality (9.1), we have

$$\mathbf{P}\{\mathbf{Z} \in \mathbb{B}\} = \mathbf{P}\{\|\mathbf{Z}\| \leq r(V)\} \leq \mathbf{P}\{\|\mathbf{Z}\| \leq \sqrt{n(P_N - \epsilon)}\} < \epsilon,$$

eventually. Thus, if the rate $R = \frac{1}{n} \log_2 M$ of transmission in bits per dimension satisfies $R \geq \frac{1}{2} \log_2 \left(1 + \frac{P_S + 2\epsilon}{P_N - \epsilon}\right)$ then the arithmetic mean of the probabilities that the receiver makes a correct decision is bounded by 2ϵ for all sufficiently large n —*a fortiori*, the probability of successful recovery is no more than 4ϵ for at least one-half of the signal vectors—and reliable communication is not possible for *any* choice of code book and decision function. Simple algebra shows that, for $0 < \epsilon \leq P_N/2$, the expression $(P_S + 2\epsilon)/(P_N - \epsilon)$ differs from P_S/P_N in no more than $4\epsilon \max\{P_S/P_N, P_N/P_S\}$ which may be made as small as desired by choosing a tiny ϵ . We are hence led to focus on the quantity

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P_S}{P_N}\right), \quad (9.4)$$

our essentially geometric analysis showing that if $R > C$ then a substantive fraction of the signal vectors cannot be recovered with high probability for *any* choice of power-limited code book and decision function.

THEOREM 1 *It is not possible to communicate reliably at any rate $R > C$ bits per dimension over a power-limited Gaussian channel.*

Having determined what cannot be done the reader may well wonder what can be. In a *tour de force* in 1948, Claude E. Shannon showed remarkably that any rate $R < C$ is indeed achievable.⁶

⁶C. E. Shannon, “A mathematical theory of communication”, *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

THEOREM 2 *It is possible to communicate reliably at any rate $R < C$ bits per dimension over a power-limited Gaussian channel.*

Shannon's theorem asserts, for any $\epsilon > 0$ and sufficiently large n , the existence of a power-limited code book of $M = 2^{nR}$ signal vectors and a decision function for which the probability of recovering any transmitted signal vector is in excess of $1 - \epsilon$.

The quantity C is called the *Gaussian channel capacity* and is a fundamental of information theory. A generation of communication engineers believed before Shannon that noise places an ineluctable limit on how well one can communicate. Shannon turned intuition on its head and made the remarkable contrarian assertion that noise places no restriction on how well one can communicate and that indeed it is possible to communicate arbitrarily reliably at any rate below capacity. Shannon's probabilistic proof, alas!, is non-constructive. The high-speed digital communication systems of the modern age are a testament both to Shannon's genius and the ingenuity of communication engineers in designing practical systems ever closer to the limits prescribed by Shannon.

While Shannon's proof of the positive statement is not so hard that I cannot present it here, it will take us afield from our theme on integration. The reader who is curious to see how we can make progress in this regard will find the approach outlined in Section XVII.2 and fleshed out in Problems XVII.10–14. Shannon's paper presents the final word with a more sophisticated analysis along the lines I've indicated. It still makes for compelling reading.

10 Convolution, symmetrisation

Sums of independent random variables that we have hitherto encountered in particular settings play an important rôle in probability. We now return to this theme from a general vantage point.

Suppose X_1, X_2, \dots is a sequence of independent random variables, each X_k possessed of marginal distribution F_k . For each n , let $S_n = X_1 + \dots + X_n$ and let G_n be the distribution of S_n . As usual, $n = 2$ provides a model for the general case. For each t , the event $\{X_1 + X_2 \leq t\}$ may be identified with the set $A_t = \{(x_1, x_2) : x_1 + x_2 \leq t\}$ of points in the plane shown shaded in Figure 8. The integral of the product measure $F_1 \otimes F_2$ over this region is best computed as an iterated integral as suggested by the figure and we obtain

$$G_2(t) = P\{X_1 + X_2 \leq t\} = E(1_{A_t}(X_1, X_2))$$

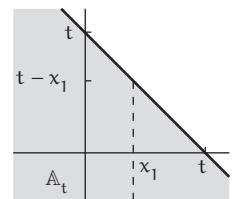


Figure 8: Convolution.

$$= \int_{-\infty}^{\infty} \left\{ \int_{(-\infty, t-x_1]} dF_2(x_2) \right\} dF_1(x_1) = \int_{-\infty}^{\infty} F_2(t - x_1) dF_1(x_1). \quad (10.1)$$

(The reader will recognise the implicit appeal to Fubini's theorem.) As the distribution of $S_2 = X_1 + X_2$ is unaffected by the order of the terms in the sum, we can interchange the rôles of the d.f.s in the integral to obtain the equivalent formulation

$$G_2(t) = \int_{-\infty}^{\infty} F_1(t - x_2) dF_2(x_2). \quad (10.1')$$

The integrals on the right of (10.1), (10.1') are deserving of special mention.

DEFINITION 1 If F_1 and F_2 are any two distribution functions, the *convolution* of F_1 and F_2 , denoted $F_1 * F_2$, is the distribution function given by

$$(F_1 * F_2)(t) = \int_{-\infty}^{\infty} F_2(t - x) dF_1(x). \quad (10.2)$$

The $*$ notation for convolution is a little overburdened—it means something slightly different depending on whether we're dealing with distributions of lattice variables (VII.2.1), or densities of absolutely continuous variables (VII.10.1), or distribution functions of arbitrary random variables as in the present context (10.1). It does not appear worthwhile, however, to take on the ballast of new notation to handle each individual case; there is little danger of confusion and the context makes clear the particular usage in force.

The relation (10.1) establishes that the convolution of distributions is always a distribution function: $F_1 * F_2$ may be identified as the d.f. of the sum of two independent random variables whose marginal d.f.s are F_1 and F_2 . While the definition (10.2) is asymmetric in the way it treats F_1 and F_2 , (10.1') establishes that the order does not matter, $F_1 * F_2 = F_2 * F_1$, that is to say, convolution of distributions is commutative. Summarising our findings we have the following important theorem which covers and extends our previous results.

THEOREM 1 *If X_1 and X_2 are independent with marginal d.f.s F_1 and F_2 , respectively, then the d.f. of the sum $S_2 = X_1 + X_2$ is given by $G_2 = F_1 * F_2 = F_2 * F_1$.*

In words, the distribution of a sum of independent random variables is the convolution of the corresponding marginals. While it is tempting to associate convolution with independence, Problems IX.9,10 show, sadly, that the implication does not hold in reverse: *if the distribution of a sum is given by a convolution of marginals, it does not imply that the constituent variables are independent.*

As integration connotes averaging, it is intuitive that the distribution of S_2 smears out the distributions of X_1 and X_2 ; *convolution is a smoothing operation.* (The reader who prefers a slightly more quantitative backing for the intuition will find it in the analysis leading to the slogan of Section XV.3.)

At need the notation allows us to segue smoothly from a view of the d.f. as a point function to the d.f. as a set function. If \mathbb{A} is any Borel set in the plane then the affine set $\mathbb{A} - x$ consists of the points y for which $y - x$ is in \mathbb{A} and the expressions

$$G_2(\mathbb{A}) = \int_{-\infty}^{\infty} F_2(\mathbb{A} - x) dF_1(x) = \int_{-\infty}^{\infty} F_1(\mathbb{A} - x) dF_2(x)$$

then all simply stand for the probability that $X_1 + X_2$ takes values in \mathbb{A} .

EXAMPLES: 1) *Arithmetic variables.* If X_1 and X_2 are positive arithmetic random variables then F_1 and F_2 have jumps at the integer values $j \geq 0$ with $p_1(j) = P\{X_1 = j\} = F_1(j) - F_1(j-)$ and $p_2(j) = P\{X_2 = j\} = F_2(j) - F_2(j-)$. The convolution (10.1) then reduces to $G_2(t) = \sum_{j \geq 0} F_2(t - j)p_1(j)$. It is clear that G_2 is discrete with points of jump also at the positive integers. If $q_2(k) = P\{S_2 = k\} = G_2(k) - G_2(k-)$ represents the jump of G_2 at k then $q_2(k) = \sum_{j \geq 0} p_2(k - j)p_1(j)$ which matches (VII.2.1).

2) *Absolutely continuous variables.* If X_1 and X_2 have densities f_1 and f_2 , respectively, then (10.1) reduces to $G_2(t) = \int_{-\infty}^{\infty} F_2(t - x)f_1(x) dx$. Differentiating with respect to t yields $g_2(t) = \int_{-\infty}^{\infty} f_2(t - x)f_1(x) dx$ for the density of S_2 , a result which matches (VII.10.1).

3) *Singular continuous variables.* Suppose $\{Z_k, k \geq 1\}$ is a sequence of Bernoulli trials corresponding to the flips of a fair coin. Then $X_1 = .Z_10Z_30Z_50\dots$ and $X_2 = .0Z_20Z_40Z_6\dots$ (both expansions representing real numbers in base 2) are both singular continuous random variables whose d.f.s are variants of the Cantor distribution. On the other hand, the random variable $S_2 = X_1 + X_2 = .Z_1Z_2Z_3Z_4\dots$ is absolutely continuous—it is uniformly distributed in the unit interval! Thus, *convolutions of singular continuous distributions can result in absolutely continuous distributions.*

4) *Mixed sums.* Suppose X_1 takes values 0 and $1/2$, each with probability $1/2$. Suppose X_2 has the uniform density $u(t)$, with corresponding d.f. $U(t)$, in the unit interval. Then (10.1) reduces to the expression $G_2(t) = \frac{1}{2}[U(t) + U(t - 1/2)]$. Differentiation shows that the sum $S_2 = X_1 + X_2$ hence has density $g_2(t) = \frac{1}{2}[u(t) + u(t - 1/2)]$ with support in the interval $(0, 3/2)$. ▶

The extension to general n is straightforward: as $S_n = S_{n-1} + X_n$ is a sum of two independent random variables, induction carries the day. *The d.f. of the sum $S_n = X_1 + \dots + X_n$ is given by the n -fold convolution $G_n = G_{n-1} * F_n = F_1 * \dots * F_n$.* Again, the distribution of the sum does not depend upon the order of the terms in the sum so that the convolutions can be performed in any order—convolution is both associative and commutative. As a matter of notation, we write F^{*n} for the n -fold convolution of a distribution F with itself; thus,

if X_1, \dots, X_n are independent with a common distribution F then the sum S_n has distribution F^{*n} .



SYMMETRISATION

No new ground is broken if we consider differences of independent random variables. First some notation.

If X has d.f. F , write \bar{F} for the d.f. of $Y = -X$. As $P\{Y \leq y\} = P\{X \geq -y\}$, it is easy to see that $\bar{F}(y) = 1 - F(-y)$ at any point of continuity of F . The fact that a d.f. must be continuous from the right allows us to finish off the specification of \bar{F} .

If X_1 and X_2 are independent with d.f.s F_1 and F_2 , respectively, it follows that the difference $X_1 - X_2 = X_1 + (-X_2)$ has d.f. $F_1 * \bar{F}_2 = \bar{F}_2 * F_1$. The case when X_1 and X_2 have a common d.f. F is of particular interest and we assume that this is the case for the rest of this section. Under this assumption, introduce the notation 0F for the d.f. of $X_1 - X_2$. We then have

$${}^0F(t) = (\bar{F} * F)(t) = \int_{\mathbb{R}} F(t-y) \bar{F}(dy).$$

A change of variable allows us to write the integral as an expectation over the d.f. F itself. Writing X for a generic variable with d.f. F , the variable $Y = -X$ has d.f. \bar{F} . We may hence express the integral on the right compactly in the form $E\{F(t-Y)\} = E\{F(t-(-X))\}$. The change of variable $x \leftarrow -y$ inside the expectation integral shifts vantage point to the distribution of X itself and we obtain

$${}^0F(t) = E(F(t+X)) = \int_{\mathbb{R}} F(t+x) F(dx).$$

This relation simplifies in the usual ways in discrete and continuous cases. If F is arithmetic with atomic probabilities $p(k)$ then 0F is also arithmetic with symmetrised atomic probabilities ${}^0p(-k) = {}^0p(k)$ given by ${}^0p(k) = \sum_{j=-\infty}^{\infty} p(k+j)p(j)$. If, on the other hand, F is absolutely continuous with density f then 0F is absolutely continuous with symmetrised density ${}^0f(-t) = {}^0f(t)$ given by ${}^0f(t) = \int_{-\infty}^{\infty} f(t+x)f(x) dx$.

The reason for considering the d.f. 0F instead of F is that 0F has a symmetry property that the original d.f. F may not possess.

DEFINITION 2 A d.f. G is *symmetric* if $G(t) = 1 - G(-t)$ at all points of continuity of G .

Now observe that $P\{X_1 - X_2 \leq t\} = P\{X_2 - X_1 \geq -t\} = P\{X_1 - X_2 \geq -t\}$ as the variables are independent and possessed of the common distribution F . It follows that 0F is symmetric and, accordingly, we say that the d.f. 0F is obtained via *symmetrisation* of F . This codifies the intuitive idea that symmetrisation properly centres the distribution at the origin. A *definition* of centring makes matters precise.

DEFINITION 3 A *median* of a random variable X (or the associated distribution) is any value M for which $P\{X \geq M\} \geq 1/2$ and $P\{X \leq M\} \geq 1/2$.

A median always exists for any random variable; in particular, we may select as median the unique value $\min\{m : P\{X \leq m\} \geq 1/2\}$. Thus, we see that 0 is a median of 0F and, furthermore, if 0F has expectation then it must also be zero.

Variations on a Theme of Integration

EXAMPLES: 5) *Symmetrised Bernoulli trials.* If a Bernoulli trial has success probability $p(1) = p$ and failure probability $p(0) = q = 1 - p$, the symmetrised distribution has atoms at -1 , 0 , and 1 with corresponding probabilities ${}^0p(-1) = {}^0p(1) = pq$ and ${}^0p(0) = 1 - 2pq$.

Observe on the other hand that the distribution with atoms of equal probability $1/2$ at -1 and $+1$ is symmetric but *does not arise out of a symmetrisation procedure*.

6) *Symmetrised binomial.* Symmetrisation of the binomial distribution $b_n(k) = \binom{n}{k} 2^{-n}$ yields the distribution with atomic probabilities

$${}^0b_n(k) = 2^{-2n} \sum_j \binom{n}{j} \binom{n}{k+j} = 2^{-2n} \sum_j \binom{n}{j} \binom{n}{n-k-j}.$$

The term on the right sums over all ways of selecting $n - k$ objects from $2n$ objects partitioned into two groups of n apiece. (This is Vandermonde's convolution.) It follows that ${}^0b_n(k) = 2^{-2n} \binom{2n}{n-k}$. The symmetrised binomial has atoms at $0, \pm 1, \dots, \pm n$.

7) *Symmetrised Poisson.* The Poisson distribution with mean α has atomic probabilities $p(k) = e^{-\alpha} \alpha^k / k!$ for $k \geq 0$. Symmetrisation of the Poisson yields a symmetric arithmetic distribution with support on all the integers, ${}^0p(-k) = {}^0p(k)$, and satisfying

$${}^0p(k) = e^{-2\alpha} \sum_{j=0}^{\infty} \frac{\alpha^{k+2j}}{j!(k+j)!} = e^{-2\alpha} I_k(2\alpha) \quad (k \geq 0)$$

where $I_k(\cdot)$ is the modified Bessel function of the first kind of order k .

8) *Symmetrised waiting times.* If $w(k) = q^k p$ connotes the waiting time till the first success in a succession of coin tosses with success probability p , then the symmetrised geometric distribution is given by ${}^0w(k) = q^{|k|} p / (1 + q)$ with support on the entire set of integers.

9) *Symmetrised uniform.* If $u(x)$ is the uniform density on the unit interval, symmetrisation yields the triangular density ${}^0u(t) = 1 - |t|$ with support in the interval $(-1, 1)$.

10) *Symmetrised exponential.* Symmetrisation of the exponential density $g(x) = \alpha e^{-\alpha x}$ ($x \geq 0$) yields the bilateral exponential ${}^0g(t) = \frac{\alpha}{2} e^{-\alpha|t|}$ with support on the entire line.

11) *Symmetrised normal.* If X_1 and X_2 are independent, standard normal random variables then $X_1 - X_2$ is normal with mean zero and variance 2. Thus, symmetrisation of the normal density ϕ results in the normal density ${}^0\phi(t) = 2^{-1/2} \phi(t2^{-1/2})$. ▶

The symmetrised distribution 0F is a better candidate for analysis than the original distribution F as its tails are better behaved than those of F . The link between the two is provided by the following symmetrisation inequalities.

THEOREM 2 (SYMMETRISATION INEQUALITIES) Suppose X_1 and X_2 are independent random variables drawn from a common distribution and suppose M is any median of these variables. Then for every $t > 0$ we have

$$\frac{1}{2} \mathbf{P}\{|X_1 - M| \geq t\} \leq \mathbf{P}\{|X_1 - X_2| \geq t\} \leq 2 \mathbf{P}\{|X_1 - M| \geq \frac{t}{2}\}.$$

PROOF: As $|X_1 - X_2| \leq |X_1 - M| + |X_2 - M|$ by the triangle inequality, for the event $\{|X_1 - X_2| \geq t\}$ to occur it is necessary that at least one of the two events $\{|X_1 - M| \geq t/2\}$

and $\{|X_2 - M| \geq t/2\}$ occur. As X_1 and X_2 have the same distribution, Boole's inequality wraps up the upper bound.

For the lower bound, observe that for the event $\{|X_1 - X_2| \geq t\}$ to occur it suffices if either the event $\{X_1 \geq M + t, X_2 \leq M\}$ occurs or the event $\{X_1 \leq M - t, X_2 \geq M\}$ occurs. (If the reader sketches the regions in the plane suggested by these inequalities all will be clear.) The latter two events are mutually exclusive and so

$$\begin{aligned} P\{|X_1 - X_2| \geq t\} &\geq P\{X_1 \geq M + t\} P\{X_2 \leq M\} + P\{X_1 \leq M - t\} P\{X_2 \geq M\} \\ &\geq \frac{1}{2} P\{X_1 - M \geq t\} + \frac{1}{2} P\{X_1 - M \leq -t\} = \frac{1}{2} P\{|X_1 - M| \geq t\} \end{aligned}$$

as X_1 and X_2 are independent and have the same distribution. ▶

A variation on this theme is a key component in the proof of uniform laws of large numbers. We shall return to these ideas in Section XVII.11.

11 Labeyrie ponders the diameter of stars

It is hard to overstate the importance of convolution in mathematics. In probability convolution appears as a mixing operation, courtesy total probability, when dealing with sums of independent variables. But it crops up in all kinds of other contexts, its ubiquity explained in part by its connection to linearity. Here is an example which is of some independent interest, an unexpected probabilistic twist adding zest to the tale.

The reader is aware of the celestial phenomenon that the moon when full appears as a crisply defined disc in the night sky, but the stars twinkle. This phenomenon, called radio scintillation, originates because atmospheric effects cause distortion on scales comparable to the stellar diameters that are seen from earth, but are negligible on the scale of visible lunar diameter. As images from space-based telescopes make abundantly clear, without the obscuring effects of the atmosphere the stars would be visible as tiny discs of illumination.

A planet-bound astronomer can, by taking two observations of a star spaced six months apart, form an estimate of its distance from earth. (The observations at opposite points in the earth's orbit around the sun form two points on a surveyor's baseline. Elementary geometry allows one to estimate the stellar distance by triangulation.) If she could observe the stellar disc without distortion then, given the stellar distance, she would be able to calculate the true diameter of the star from the diameter of its observed image. The question then arises whether it may be possible to cancel out the blurring effects of the atmosphere and thence estimate true stellar diameters. The French astronomer Antoine Labeyrie showed how in 1970 by a very simple technique called speckle interferometry.⁷ It is simplest to illustrate Labeyrie's idea in a one-dimensional setting.

⁷A. Labeyrie, "Attainment of diffraction limited resolution in large telescopes by Fourier analysing speckle patterns in star images", *Astronomy and Astrophysics*, vol. 6, pp. 85–87, 1970.

ONE-DIMENSIONAL STARS

To begin, consider a bright point source (an impulse!) of amplitude s imaged at the origin of the line. The blurring effect of the atmosphere is manifest in a spreading out of the source so that the recorded image has an amplitude spread on the line given by a function $sK(x)$ where the kernel K represents the “impulse response” of the atmosphere. If the point source were located at the point x_0 then, assuming that the atmospheric distortion is spatially invariant, an assumption that is not unreasonable on small spatial scales, the observed response would be $sK(x - x_0)$ corresponding merely to a shift of origin to the point x_0 . The reader will recognise the beginnings of a translation-invariant, linear model for the atmospheric operator. Building on this intuition, if the source has a spatial amplitude profile $s(x)$ then the recorded blurred profile r is given by the convolution

$$r(x) = \int_{-\infty}^{\infty} s(y)K(x - y) dy = (K * s)(x). \quad (11.1)$$

Figure 9 shows a sample kernel. As convolution is a smoothing operation, with-

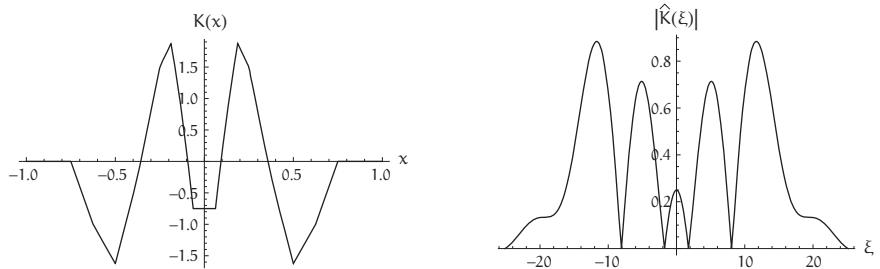


Figure 9: A stylised one-dimensional blurring kernel $K(x)$ and the modulus of its spectrum $|\hat{K}(\xi)|$. The kernel was selected purely for illustrative purposes and makes no pretence at realism.

out knowing the atmospheric kernel K it seems hopeless to recover the star profile from the blurred response.

Before we give it up as a hopeless job we should examine the particular structure of the problem in more detail. If we assume, as is reasonable, that the star has a uniform brightness profile and, viewed from a (one-dimensional) planetary perspective, a width (diameter) D centred at x_0 then, on a suitably normalised intensity scale, $s(x) = \text{rect}\left(\frac{x-x_0}{D}\right)$ in the notation introduced in Section VI.2. Its Fourier transform [see Example VI.2.1]

$$\hat{s}(\xi) = \widehat{\text{Rect}}(D\xi) e^{-i\xi x_0} = D \text{sinc}\left(\frac{D\xi}{2}\right) e^{-i\xi x_0}$$

hence exhibits regularly spaced zeros at the points $\xi = 2n\pi/D$ for integers $n \neq 0$. The visible (or imaged) star diameter D is hence determined by the zeros of the star spectrum $\hat{s}(\xi)$.

This observation suggests that we look at the spectrum of the blurred profile. By taking Fourier transforms of both sides of (11.1), we see that $\hat{r}(\xi) = \hat{s}(\xi)\hat{K}(\xi)$ by the convolution property of the transform. If we now examine the zeros of the spectrum we come up against the inconvenient fact that the zeros of the blurring kernel spectrum $\hat{K}(\xi)$ appear intermixed with the zeros of the star spectrum $\hat{s}(\xi)$. The blurring kernel K is, of course, not known to us and it appears difficult to identify which are the zeros of \hat{K} and which of \hat{s} without knowing K . And again we appear to be at an impasse. Figure 10 illustrates the blurring of the star boundary for the stylised kernel shown in Figure 9.

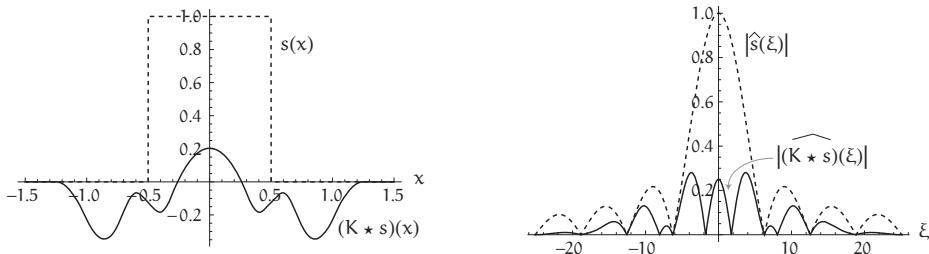


Figure 10: A rectangular star intensity profile $s(x)$ and the corresponding blurred profile $r(x) = (K * s)(x)$ are shown on the left. As we see, the blurred image of the star obscures vital features and, in particular, the star's diameter. The modulus of the spectra of the star intensity profile and that of the blurred image, $|\hat{s}(\xi)|$ and $|(K * s)(\xi)|$, respectively, are shown on the right. The extraneous zeros from the impulse response of the atmosphere show up clearly.

Labeyrie's key observation was that the blurring kernel K is not only random, but that it varies randomly in time. Thus, more accurately, $K(x) = K_t(x)$ is the kernel at time t , yielding the corresponding stellar image $r_t(x) = (K_t * s)(x)$. Consider now the effects of averaging suitably centred observations to form a composite image

$$r(x) = \frac{1}{n} \sum_{j=1}^n r_{t_j}(x) = \frac{1}{n} \sum_{j=1}^n (K_{t_j} * s)(x).$$

If we set $\bar{K}(x) = \frac{1}{n} \sum_{j=1}^n K_{t_j}(x)$, linearity of convolution shows then that $r(x) = (\bar{K} * s)(x)$. It follows that the composite image spectrum is given by

$$\hat{r}(\xi) = \hat{s}(\xi)\hat{\bar{K}}(\xi) = \hat{s}(\xi) \cdot \frac{1}{n} \sum_{j=1}^n \hat{K}_{t_j}(\xi)$$

by linearity of the transform. Now we haven't actually specified a probability model for the random process $K_t(x)$ determining the kernel. And we won't. For our purposes, it suffices to observe that, for any reasonable probability model,

the spectrum of the random kernel should have a finite number of zeros, the locations of which are random, varying independently from realisation to realisation. As it is very unlikely that a zero of the kernel spectrum recurs across different realisations, summing the blurring spectra across distinct points in time hence has the effect of eliminating the individual zeros of the kernel spectra. Thus, with just a few observations, zeros are essentially all eliminated from the composite spectrum $\widehat{K}(\xi)$ except, possibly, for some stray accidental cancellation of terms. The zeros of the composite image spectrum \widehat{r} are hence essentially only those of the star spectrum \widehat{s} , enabling us to estimate the visible stellar diameter D quite accurately.

The reader should note that we have not attempted to cancel out (or, in the jargon, *equalise*) the effects of the atmosphere to recover the original stellar image [see the discussion at the end of Section X.7]. That is a much more complex task as it will require methods of estimating a randomly varying blurring kernel function $K_t(x)$. But our task is not to reconstruct s per se but merely to determine its diameter. And the composite blurred image r , while it looks nothing like s , does splendidly for the task as the zeros of its spectrum coincide with those of s .

TWO-DIMENSIONAL STARS

Now that we understand how to proceed in the one-dimensional case, the two-dimensional images of real interest are easy to handle. The raw stellar image (at the image plane of the camera) is a two-dimensional function $s(x, y)$. By modelling the atmosphere's blurring effect as a linear, translation-invariant operator, the response at the image plane to a point source at the origin is the impulse response $K(x, y)$. Accordingly, the blurred image that is recorded is given by a convolution integral in two dimensions,

$$r(x, y) = (K * s)(x, y) = \iint_{\mathbb{R} \times \mathbb{R}} s(\alpha, \beta) K(x - \alpha, y - \beta) d\alpha d\beta.$$

As before, we model $K = K_t$ as a random spatial process so that $r_t(x, y) = (K_t * s)(x, y)$ is the recorded image (properly centred) at time t . By averaging over several observations we can build up the composite image

$$r(x, y) = \frac{1}{n} \sum_{j=1}^n r_{t_j}(x, y) = \frac{1}{n} \sum_{j=1}^n (K_{t_j} * s)(x, y) = (\bar{K} * s)(x, y)$$

where $\bar{K} = \frac{1}{n} \sum_{j=1}^n K_{t_j}$ is the average of the atmospheric effects at the observation points. Of course, this function is very far removed from s but, proceeding as for stars in one dimension, we are tempted to take Fourier transforms.

By analogy with transforms in one dimension, we define the spectrum or Fourier transform of an integrable function $f(x, y)$ in two dimensions by

$$\widehat{f}(\xi, \eta) = \iint_{\mathbb{R} \times \mathbb{R}} f(x, y) e^{-i(\xi x + \eta y)} dx dy.$$

The reader should verify that the convolution property of the transform carries through unscathed and we obtain

$$\widehat{r}(\xi, \eta) = \widehat{s}(\xi, \eta) \widehat{\mathcal{K}}(\xi, \eta) = \widehat{s}(\xi, \eta) \cdot \frac{1}{n} \sum_{j=1}^n \widehat{K}_{t_j}(\xi, \eta).$$

As in the one-dimensional case, we anticipate that the random zeros of the functions $\widehat{K}_{t_1}(\xi, \eta), \dots, \widehat{K}_{t_n}(\xi, \eta)$ will be erased by averaging so that the residual zeros of $\widehat{r}(\xi, \eta)$, if any, must be solely due to $\widehat{s}(\xi, \eta)$.

The picky reader may point out that this is all very well but we haven't actually shown that the spectrum $\widehat{s}(\xi, \eta)$ actually has any zeros. Now it is certainly reasonable to model the star as having uniform brightness in some disc. Accordingly, we begin with a consideration of the unit disc function $\text{disc}(x, y)$ which takes value 1 inside the unit disc $\{(x, y) : x^2 + y^2 < 1\}$ and takes value zero outside. In view of the angular symmetry that is implicit, in order to compute the Fourier transform of $\text{disc}(x, y)$ it is natural to shift to polar coordinates. Setting $\rho(\xi, \eta) = \sqrt{\xi^2 + \eta^2}$, we may write

$$\xi x + \eta y = \rho(\xi, \eta) \left(\frac{\xi}{\rho(\xi, \eta)} x + \frac{\eta}{\rho(\xi, \eta)} y \right) = \rho(\xi, \eta) r \cos \theta$$

where, as the pair (x, y) sweeps through the unit disc, r varies from 0 to 1 and θ varies from 0 to 2π . We hence obtain

$$\widehat{\text{disc}}(\xi, \eta) = \iint_{x^2 + y^2 < 1} e^{-i(\xi x + \eta y)} dx dy = \int_0^1 \int_0^{2\pi} e^{-i\rho(\xi, \eta)r \cos \theta} r d\theta dr.$$

The integrals on the right are expressed most compactly in terms of Bessel functions of the first kind. These may be expressed recursively via

$$\begin{aligned} J_0(z) &= \frac{1}{2\pi} \int_0^{2\pi} e^{-iz \cos \theta} d\theta, \\ z^n J_n(z) &= \int_0^z u^n J_{n-1}(u) du \quad (n \geq 1). \end{aligned}$$

Specialising to the case $n = 1$, we hence obtain

$$\begin{aligned} \widehat{\text{disc}}(\xi, \eta) &= 2\pi \int_0^1 r J_0(\rho(\xi, \eta)r) dr \stackrel{u \leftarrow \rho(\xi, \eta)r}{=} \frac{2\pi}{\rho(\xi, \eta)^2} \int_0^{\rho(\xi, \eta)} u J_0(u) du \\ &= \frac{2\pi}{\rho(\xi, \eta)^2} \cdot \rho(\xi, \eta) J_1(\rho(\xi, \eta)) = 2\pi \frac{J_1(\rho(\xi, \eta))}{\rho(\xi, \eta)}. \end{aligned}$$

The spectrum $\widehat{\text{disc}}(\xi, \eta)$ hence exhibits angular symmetry, its zeros determined in the radial direction by the characteristic zeros of the Bessel function $J_1(z)$ shown in Figure 11.

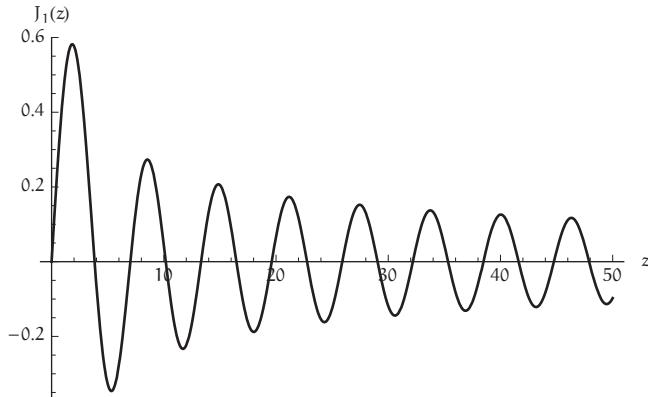


Figure 11: The graph of the Bessel function of the first kind of order one.

Let us consider now a star of uniform imaged brightness in some disc of diameter D . At the image plane of the camera we may represent a unit-intensity stellar disc of diameter D centred at coordinates (x_0, y_0) by $s(x, y) = \text{disc}\left(\frac{x-x_0}{D}, \frac{y-y_0}{D}\right)$. The shift and scale properties of the transform then immediately yield

$$\widehat{s}(\xi, \eta) = D^2 \widehat{\text{disc}}(D\xi, D\eta) e^{-i(\xi x_0 + \eta y_0)}.$$

As the zeros of Labeyrie's composite spectrum $\widehat{r}(\xi, \eta)$ coincide (with high probability) with the zeros of $\widehat{s}(\xi, \eta)$, and these in turn are inherited by scaling by D from the characteristic zeros of the spectrum of the function $\text{disc}(x, y)$, the zeros of \widehat{r} determine the diameter of the star.

Labeyrie's method has to date been put to use to measure the diameters of several dozens of stars. And while Hubble and its descendants offer scope for unprecedented resolution in astronomical imagery, planet-based astronomy continues to be an important part of the puzzle, in no small part because of the huge expenses involved in space-based astronomy.

12 Problems

1. Suppose X is uniformly distributed in the interval $[-1, 1]$. Let $Y = |X|$ and $Z = X/|X|$ (set $Z = 0$ if $X = 0$). Are Y and Z uncorrelated? Are they independent?
2. Suppose X and Y are Bernoulli variables (success probabilities not necessarily the same). If $\text{Cov}(X, Y) = 0$ determine whether X and Y are independent.
3. Suppose X and Y are independent, standard normal random variables. Determine $E(\min\{|X|, |Y|\}/\max\{|X|, |Y|\})$. [Partition the (x, y) -plane into two regions depending on whether x is larger or y is larger and integrate separately over these regions.]

4. Order statistics for the uniform. Suppose X_1, X_2, \dots, X_n are independent random variables uniformly distributed in the unit interval; rearranging them in order of increasing size $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ yields their order statistics. Consider the interval lengths $L_k = X_{(k)} - X_{(k-1)}$ of Section IX.1. As $X_{(k)} = L_1 + \dots + L_k$, additivity of expectation provides a simple path to determining $E(X_{(k)})$.

5. Continuation. Now determine $\text{Cov}(X_{(j)}, X_{(k)})$.

6. The sample variance. Suppose X_1, \dots, X_n are independent variables with common mean μ and variance σ^2 . Let $\hat{M} = \frac{1}{n}(X_1 + \dots + X_n)$ and $\hat{\Sigma}^2 = \frac{1}{n-1}[(X_1 - \hat{M})^2 + \dots + (X_n - \hat{M})^2]$ denote the sample mean and sample variance, respectively. Show that $E(\hat{\Sigma}^2) = \sigma^2$. In the language of statistics, the sample variance is an *unbiased estimator* of the population variance.

7. Length of random chains. This is a two-dimensional analogue of the length of random flights in \mathbb{R}^3 . A chain in the (x, y) -plane consists of n links of unit length. Each succeeding link forms an angle $\pm\alpha$ with the last link where α is a fixed positive constant and the two choices have equal probability $1/2$ with successive angles mutually independent. The angle between the j th link and the positive x -axis is a random variable Θ_{j-1} and we may recursively specify $\Theta_0 = 0$ and $\Theta_j = \Theta_{j-1} + \alpha X_j$ where X_1, \dots, X_{n-1} denote independent variables taking values -1 and $+1$ only, each with probability $1/2$. (There is no harm in assuming that the first link is aligned with the x -axis.) By induction, for $m < n$ prove: (a) $E(\cos \Theta_n) = \cos(\alpha)^n$ and $E(\sin \Theta_n) = 0$; (b) $E(\cos(\Theta_m) \cos(\Theta_n)) = \cos(\alpha)^{n-m} E(\cos(\Theta_m)^2)$; and (c) $E(\sin(\Theta_m) \sin(\Theta_n)) = \cos(\alpha)^{n-m} E(\sin(\Theta_m)^2)$.

8. Continuation. Let L_n denote the distance from the end of the chain to its starting point. Show that

$$E(L_n^2) = n \cdot \frac{1 + \cos \alpha}{1 - \cos \alpha} - 2 \cos(\alpha) \cdot \frac{1 - \cos(\alpha)^n}{(1 - \cos \alpha)^2}.$$

[Hint: Express L_n^2 by projections of the links along the two axes and set up an induction by evaluating $E(L_n^2 - L_{n-1}^2)$ explicitly.]

9. Another cautionary example. Show that the iterated integrals

$$\int_{-1}^1 \left\{ \int_{-1}^1 \frac{xy}{(x^2 + y^2)^2} dy \right\} dx \quad \text{and} \quad \int_{-1}^1 \left\{ \int_{-1}^1 \frac{xy}{(x^2 + y^2)^2} dx \right\} dy$$

both exist and are equal but that the double integral (with respect to Lebesgue measure on the plane)

$$\iint_{[-1,1] \times [-1,1]} \frac{xy}{(x^2 + y^2)^2} dxdy$$

fails to exist. This speaks again to the necessity of integrability in Fubini's theorem.

10. Suppose F is any distribution function. Consider what Fubini's theorem says about $\int_{\mathbb{R}} (F(x+a) - F(x)) dx$.

11. Moments. Suppose X is a positive random variable with d.f. $F(x)$. Show that

$$E(X^\nu) = \nu \int_0^\infty x^{\nu-1} [1 - F(x)] dx$$

for any $\nu > 0$ (in the sense that if one converges then so does the other).



12. *Outer integrals, again.* Fubini's theorem fails for outer integrals (see Problem XIII.24 for the definition). Why?

13. Suppose F is a continuous d.f. Show that $\int_{-\infty}^{\infty} F(x) dF(x) = \frac{1}{2}$ from first principles by the definition of the Lebesgue integral. Now redo the problem by first determining the distribution of the random variable $F(X)$ and interpreting it.

14. *On medians.* Revisit the idea of a median from Definition 10.3. (a) Show that a median always exists. (b) Is it always unique? (c) Suppose M is a median of X . If \mathbb{I} is a closed interval such that $P\{X \in \mathbb{I}\} \geq 1/2$, show that $M \in \mathbb{I}$.

15. *Continuation.* When the variance exists, show that $|M - E(X)| \leq \sqrt{2 \text{Var}(X)}$.

16. *Absolute error prediction.* If M is a median of X and X is integrable, show that $E(|X - M|) \leq E(|X - a|)$ for all real a .

17. *Squared error prediction.* If X has a finite second moment and $\mu = E(X)$, show that $E(|X - \mu|^2) \leq E(|X - a|^2)$ for all real a .

18. *Least squares prediction.* Suppose X_0, X_1, \dots, X_n have zero mean, unit variance, and a common covariance $\rho = \text{Cov}(X_j, X_k)$. Explicitly find the *best linear predictor* $\hat{X}_0^{(2)}$ of the variable X_0 based on X_1 and X_2 . That is to say, find the linear function $a_1 X_1 + a_2 X_2$ such that $E[(a_1 X_1 + a_2 X_2) - X_0]^2$ is minimised.

19. *Continuation.* Generalise by finding the *best linear predictor* $\hat{X}_0^{(n)}$ of X_0 based on X_1, \dots, X_n ; that is, find the linear function $\sum_{k=1}^n a_k X_k$ such that $E[(\sum_k a_k X_k - X_0)^2]$ is minimised. [Hint: If $A(\alpha)$ is a non-singular square matrix of order n with diagonal elements 1 and off-diagonal elements all equal to a constant α , then its inverse $[A(\alpha)]^{-1}$ must be of the form $\beta_n A(\gamma_n)$ for suitable scalars β_n and γ_n determined by α .]

20. *Energy minima on a cube.* Consider the set of vertices $v = (v_1, \dots, v_n) \in \{0, 1\}^n$ of the n -dimensional cube. To each $v \in \{0, 1\}^n$ we independently assign a standard normal random variable H_v which we call the *altitude* of v . We say that a given vertex $v \in \{0, 1\}^n$ is in a *valley* if the altitude of v is less than that of each of its neighbours (the n vertices which differ from v in exactly one component). By conditioning on the value of H_v evaluate the probability that v is in a valley. Is there a direct path to the answer? This setting occurs in Ising spin glass models in statistical physics and in models of neural computation in which contexts H_v is sometimes, possible negative values notwithstanding, identified as the "energy" of v .⁸

21. *Continuation.* Let the random variable N denote the number of valleys in the space $\{0, 1\}^n$. Determine $E(N)$ and $\text{Var}(N)$. [Hint: Express N as a sum of indicators.]

22. Show that the convolution of three singular continuous distributions can result in an absolutely continuous distribution.

23. *Spherical coordinates.* The system of equations

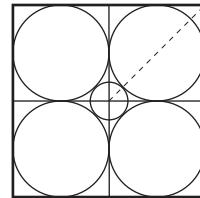
$$\begin{aligned} x_1 &= r \cos(\theta_1) \\ x_2 &= r \sin(\theta_1) \cos(\theta_2) \end{aligned}$$

$$\begin{aligned} &\dots \\ x_{n-1} &= r \sin(\theta_1) \sin(\theta_2) \cdots \sin(\theta_{n-2}) \cos(\theta_{n-1}) \\ x_n &= r \sin(\theta_1) \sin(\theta_2) \cdots \sin(\theta_{n-2}) \sin(\theta_{n-1}) \end{aligned}$$

⁸S. S. Venkatesh and P. Baldi, "Random interconnections in higher-order neural networks", *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 274–282, 1993.

describes the natural generalisation of polar transformations to n dimensions. Here the radial parameter r is positive, the angular parameters $\theta_1, \dots, \theta_{n-2}$ vary from $-\pi/2$ to $\pi/2$, and the final angular parameter θ_{n-1} varies from $-\pi$ to π . Show by induction or otherwise that the Jacobian of the transformation from Cartesian to spherical coordinates is $r^{n-1} \sin(\theta_1)^{n-2} \sin(\theta_2)^{n-3} \cdots \sin(\theta_{n-2})$. Hence evaluate $V_n = \int_{B^n(1)} dx$.

24. Surprises in high dimensions. Let C^n be the set of points $x = (x_1, \dots, x_n)$ with $|x_i| \leq 1$ for each i . In words, C^n is the axis-parallel cube of side 2 centred at the origin in \mathbb{R}^n . For each coordinate axis i , consider the planes $x_i = 0$ and $x_i = \pm 1$ orthogonal to it. These planes partition C^n into 2^n orthants, each of which is a cube of side 1. In each orthant embed an n -dimensional ball of radius 1 centred in the orthant and, finally, embed a ball centred at the origin and whose radius r_n is such that it just touches each of the 2^n balls of radius 1 in the orthants surrounding it. Determine the radius r_n of the small ball in the centre and comment on your findings for $n > 9$.



25. Continuation. Show that for $n \geq 1206$ the volume of the cube C^n is dominated by the volume of the small ball at its centre and, as $n \rightarrow \infty$, the ratio of the volume of the cube to that of the ball goes to zero.

26. Binomial distribution. Let $B_n(x; p) = \sum_{k \leq x} b_n(k; p)$ denote the d.f. of the binomial distribution. For integer k show that $B_{n+1}(k; p) = B_n(k; p) - pb_n(k; p)$.

27. Continuation, expected deviation from the mean. Let S_n denote the accumulated number of successes in n tosses of a coin with success probability p (and failure probability $q = 1 - p$). Let v denote the unique integer such that $np < v \leq np + 1$. Show that $E(|S_n - np|) = 2vqb_n(v; p)$. Conclude for large n that the expected deviation of the binomial from its mean is of the order of \sqrt{n} .

28. Optimal decisions. A code book of M signal vectors $\{s_1, \dots, s_M\}$ is given. Signal vector transmissions are garbled by additive Gaussian noise $Z = (Z_1, \dots, Z_n)$ where the components Z_k are independent and have the common $\mathcal{N}(0, P_N)$ distribution. Let $\mathbb{I}_1, \dots, \mathbb{I}_M$ be the decision regions induced by the receiver's decision function $D: \mathbb{R}^n \rightarrow \{s_1, \dots, s_M\}$. Writing $f(\cdot)$ for the density of Z , show that the receiver maximises the average probability of a correct decision, $\frac{1}{M} \sum_{j=1}^M \int_{\mathbb{I}_j} f(x - s_j) dx$, via the *minimum distance decision rule* $D^*: x \mapsto s_j$ which maps x into s_j if $\|x - s_j\| < \|x - s_k\|$ for all $k \neq j$. (Ties may be broken in any convenient fashion, say, lexicographic.)

29. Continuation. If signal vector s_j is chosen for transmission with probability P_j (where $\{P_1, \dots, P_M\}$ forms a discrete probability distribution), determine a decision rule that maximises the expected probability of a correct decision, $\sum_{j=1}^M P_j \int_{\mathbb{I}_j} f(x - s_j) dx$.

A smorgasbord of basic inequalities: these problems are of a more theoretical character.

30. Mengoli's inequality. If $x > 1$ show that $\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1} > \frac{3}{x}$. [A direct algebraic proof is easy and has the virtue of following in Pietro Mengoli's footsteps. Jensen's inequality makes it trite.] Writing $H_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}$ for the partial sums of the harmonic series, show hence that $H_{3n+1} \geq 1 + H_n$ and hence that $H_{3^k+3^{k-1}+\cdots+1} \geq k+1$. Conclude that the harmonic series diverges, $\lim_n H_n = \infty$.

31. The log sum inequality. Suppose $\{a_k\}$ and $\{b_k\}$ are positive sequences with convergent series $A = \sum_k a_k$ and $B = \sum_k b_k$. Show that $\sum_k a_k \log \frac{a_k}{b_k} \geq A \log \frac{A}{B}$ with equality holding if, and only if, there is a constant λ with $a_k = \lambda b_k$ for each k .

32. *An easy application of Cauchy–Schwarz:* $E(XYZ)^4 \leq E(X^4)E(Y^4)E(Z^2)^2$.

33. *Reciprocal bound.* Suppose X is positive. Show that $E(X) \cdot E(X^{-1}) \geq 1$. [The splitting trick $1 = X^{1/2} \cdot X^{-1/2}$ sets up Cauchy–Schwarz.] Verify that $E(X^{-1}) \geq 1/E(X)$ directly by Jensen’s inequality.

34. *Kantorovich’s inequality.* Suppose X is positive, $m \leq X \leq M$. Let $a = (m + M)/2$ and $g = \sqrt{mM}$. Show that $E(X) \cdot E(X^{-1}) \leq a^2/g^2$. This has been used to estimate the rate of convergence of steepest descent in numerical analysis. [Hint: Argue that the bound is unaffected by scaling X by a positive constant so that we may as well suppose that $\mu \leq X \leq \mu^{-1}$ for some μ . With $\alpha = \frac{1}{2}(\mu + \mu^{-1})$, show then that $E(X) + E(X^{-1}) \leq 2\alpha$. The result now sets up nicely for the inequality of arithmetic and geometric means.]

35. *The Harker–Kasper inequality.* If $f(t) = \cos(tx)$ then we have the elementary trigonometric inequality $f(t)^2 = \frac{1}{2}(1 + f(2t))$. Show that $g(t) = E \cos(tx)$ satisfies $g(t)^2 \leq \frac{1}{2}(1 + g(2t))$. This has profound consequences in crystallography. [The 1-trick $\cos(tx) = \cos(tx) \cdot 1$ can be combined with Cauchy–Schwarz to devastating effect.]

36. *Monotonicity of the L^p -norm.* If $1 \leq s < t$ show that $\|X\|_s \leq \|X\|_t$. [Identify $|Y|^s = |X|$ and use Corollary 4.1 with $p = t/s$.]

37. *Bounds on the L^p -norm.* Set $c_p = 1$ if $0 \leq p \leq 1$ and $c_p = 2^{p-1}$ if $p > 1$. If X and Y are positive variables show that $E((X+Y)^p) \leq c_p(E(X^p) + E(Y^p))$. [The reader may wish to begin by exhibiting the simpler bound with c_p replaced by 2^p .]

38. *Bounds on moments of sums.* Let $S_n = X_1 + \dots + X_n$. If $p > 1$ show that

$$\left\| \frac{1}{n} S_n \right\|_p^p \leq \frac{1}{n} \min \{ \|X_1\|_p^p + \dots + \|X_n\|_p^p, \|X_1\|_p + \dots + \|X_n\|_p \}.$$

39. Define $\varphi(p) = \log(\|X\|_p^p) = \log E(|X|^p)$ for $p \geq 1$. Show that φ is convex.

40. *Another metric.* Show that the functional $d(X, Y) = E\left(\frac{|X-Y|}{1+|X-Y|}\right)$ satisfies the triangle inequality $d(X, Y) \leq d(X, Z) + d(Z, Y)$. Thus, $d(\cdot, \cdot)$ is a metric in the space of random variables (provided that we identify equivalent random variables).

41. *A correlation inequality.* Endow $\{0, 1\}^n$ with the partial order $x \leq y$ if $x_j \leq y_j$ for each j . Say that $f: \{0, 1\}^n \rightarrow \mathbb{R}$ is *increasing* if $f(x) \leq f(y)$ whenever $x \leq y$, *decreasing* if $f(x) \geq f(y)$ whenever $x \leq y$. Suppose X_1, \dots, X_n are independent Bernoulli trials, not necessarily identical, $X_j \sim \text{Bernoulli}(p_j)$. Show that if f and g are either both increasing or both decreasing on $\{0, 1\}^n$ then $\text{Cov}(f(X)g(X)) \geq 0$; the inequality is reversed if one is increasing and the other decreasing.⁹ [Hint: The bulk of the proof is covered by the case $n = 1$ which requires only algebra. For the induction step condition on X_{n+1} .]

42. *Fisher information.* Suppose $f(x; \theta)$ is a family of densities where θ is a real parameter in some parameter class Θ . For a given θ write E_θ for expectation with respect to $f(\cdot; \theta)$. Now suppose θ is fixed and X is a random variable drawn according to the density $f(\cdot; \theta)$. Let $V(X) = \frac{\partial}{\partial \theta} \log f(X; \theta)$. Assuming the family $\{f(\cdot; \theta), \theta \in \Theta\}$ is sufficiently regular (see Theorem XXI.2.4), show that $E V(X) = \frac{d}{d\theta} \int_{\mathbb{R}} f(x; \theta) dx = 0$. The

⁹This is the simplest case of the *FKG inequality* named after C. M. Fortuin, P. W. Kasteleyn, and J. Ginibre, “Correlation inequalities on some partially ordered sets”, *Communications of Mathematical Physics*, vol. 22, pp. 89–103, 1971. A rich theory has since developed with applications in a variety of areas in statistical physics, probabilistic combinatorics, and algorithmic computer science.

quantity $I(\theta) = \text{Var } V(X)$ is called the Fisher information; it is a measure of the information that a random variable X drawn according to the density $f(x; \theta)$ carries about θ . Writing $f_\theta = \frac{\partial}{\partial \theta} f$, show that $I(\theta) = \int_{\mathbb{R}} \left(\frac{f_\theta(x; \theta)}{f(x; \theta)} \right)^2 f(x; \theta) dx = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right]$.

43. *Continuation, the Cramér–Rao lower bound.* A function $T(x)$ is called an *unbiased estimator* of θ if $E_\theta T(X) = \theta$ for every $\theta \in \Theta$. If $T(X)$ is an unbiased estimator of θ , show that $\text{Cov}(T(X), V(X)) = \int_{\mathbb{R}} T(x) f_\theta(x; \theta) dx = 1$. Hence, by Cauchy–Schwarz, prove the important Cramér–Rao inequality $\text{Var } T(X) \geq 1/I(\theta)$ which is of a basic importance in mathematical statistics.

44. *Mutual information and a test for independence.* Suppose the pair of random variables (X_1, X_2) has density $f(x_1, x_2)$ with corresponding marginals $f_1(x_1)$ and $f_2(x_2)$. The *mutual information* between X_1 and X_2 is denoted $I(X_1; X_2)$ and defined by¹⁰

$$I(X_1; X_2) = \iint_{\mathbb{R}^2} \log \left(\frac{f(x_1, x_2)}{f_1(x_1) f_2(x_2)} \right) f(x_1, x_2) dx_1 dx_2$$

provided the integral converges. Show that $I(X_1; X_2) \geq 0$ with equality if, and only if, X_1 and X_2 are independent. [Hint: Extend Kullback–Leibler divergence to densities.]

45. *Binomial exponents.* Suppose $0 < a, p < 1$, and $D(a, p)$ is the Kullback–Leibler divergence between the Bernoulli distributions with success probabilities a and p , respectively. With $b_n(k; p) = \binom{n}{k} p^k (1-p)^{n-k}$ representing the binomial distribution as usual, show that $-\frac{1}{n} \log b_n(\lfloor an \rfloor; p) \rightarrow D(a, p)$ as $n \rightarrow \infty$. Conclude anew that the Kullback–Leibler divergence is positive and $D(a, p) = 0$ if, and only if, $a = p$.

The geometry of L^2 -spaces is most evident in orthogonal projections. This geometrical vantage point provides the most elegant perspective for the abstract conditional expectations defined in the preamble to Problem XIII.13. In the following problems, the reader should assume that the space L^2 is complete (or better, read the small print in Section 4).

46. If X and Y are in $L^2(\Omega, \mathcal{F}, \mathbf{P})$, prove the *parallelogram law* $2\|X\|_2^2 + 2\|Y\|_2^2 = \|X + Y\|_2^2 + \|X - Y\|_2^2$ and explain the name.

47. *Orthogonal projections.* Suppose \mathcal{G} is a sub- σ -algebra of \mathcal{F} . For each $X \in L^2(\Omega, \mathcal{F}, \mathbf{P})$, define $\Delta(X; \mathcal{G}) = \inf \{ \|Z - X\|_2 : Z \in L^2(\Omega, \mathcal{G}, \mathbf{P}) \}$. Show that there exists a \mathcal{G} -measurable function $Y \in L^2(\Omega, \mathcal{G}, \mathbf{P})$ such that $\|Y - X\| = \Delta(X; \mathcal{G})$. We call any \mathcal{G} -measurable function with this property an *orthogonal projection* of X onto the subspace $L^2(\Omega, \mathcal{G}, \mathbf{P})$; the reason for the nomenclature will become apparent in the following problem. [Hint: Let $\{Y_n, n \geq 1\}$ be a sequence of square-integrable, \mathcal{G} -measurable functions such that $\|Y_n - X\|_2 \rightarrow \Delta(X; \mathcal{G})$. Argue that $\|Y_r - Y_s\|_2^2 = 2\|Y_r - X\|_2^2 + 2\|Y_s - X\|_2^2 - 4\left\| \frac{1}{2}(Y_r + Y_s) - X \right\|_2^2 \rightarrow 0$ as $r, s \rightarrow \infty$, and hence that $\{Y_n, n \geq 1\}$ is a Cauchy sequence.]

¹⁰The concept of mutual information was fleshed out by Claude E. Shannon in his epochal paper of 1948 in which was laid the foundation of a sweeping mathematical theory of communication [C. E. Shannon, *op. cit.*]. Shannon's notation for mutual information is slightly misleading in that it suggests that $I(X_1; X_2)$ is a function of the random variables X_1 and X_2 instead of, as is really the case, a function of the *law* (or distribution) $\mathcal{L}(X_1, X_2)$ of the pair. A more pedantically accurate notation along the lines of $I(\mathcal{L}(X_1; X_2))$ would not only be unnecessarily cumbersome but would be scuppered in any case by weight of tradition.

48. Pythagoras's theorem. Say that square-integrable functions U and V are *orthogonal*, denoted $U \perp V$, if $\langle U, V \rangle = E(UV) = 0$. Let Y be a \mathcal{G} -measurable function as in the previous problem, that is, $\|Y - X\|_2 = \Delta(X; \mathcal{G})$. Show that $Z \perp (Y - X)$ for every $Z \in L^2(\Omega, \mathcal{G}, P)$, and hence demonstrate the validity of Pythagoras's theorem $\|X\|_2^2 = \|Y\|_2^2 + \|Y - X\|_2^2$. [Hint: For every real t we have $\Delta(X; \mathcal{G})^2 \leq \|(Y + tZ) - X\|_2^2$. Argue that this leads to a contradiction for t of small modulus unless $Z \perp (X - Y)$.]

49. Orthogonal projections are identified only a.e. If $Z \perp (\tilde{Y} - X)$ and $Z \perp (Y - X)$ for every $Z \in L^2(\Omega, \mathcal{G}, P)$, then $\tilde{Y}(\omega) = Y(\omega)$ a.e. (or, equivalently, $\|\tilde{Y} - Y\|_2 = 0$).

50. Conditional expectations for square-integrable variables. Suppose \mathcal{G} is a sub- σ -algebra of \mathcal{F} . For each $X \in L^2(\Omega, \mathcal{F}, P)$, let Y denote an orthogonal projection of X onto $L^2(\Omega, \mathcal{G}, P)$. Show that $E(X1_A) = E(Y1_A)$ for every $A \in \mathcal{G}$. Hence conclude that $Y = E(X | \mathcal{G})$ is a version of a conditional expectation of X given \mathcal{G} . [A preview was provided in Theorem VII.7.1.]

51. Continuation, conditional expectations for integrable variables. If X is integrable, show that there exists a \mathcal{G} -measurable random variable $E(X | \mathcal{G})$ such that $E(X1_A) = E(E(X | \mathcal{G})1_A)$ for every $A \in \mathcal{G}$. [Hint: It suffices to prove the result for positive variables X (for the general case set $X = X^+ - X^-$). Select bounded variables $0 \leq X_n \uparrow X$ and let $Y_n = E(X_n | \mathcal{G})$ be an orthogonal projection of X_n onto $L^2(\Omega, \mathcal{G}, P)$. Show that $0 \leq Y_n \uparrow$ a.e. and finish up via monotone convergence.]

52. Conditional Jensen. As before, X is measurable with respect to \mathcal{F} and \mathcal{G} is a sub- σ -algebra of \mathcal{F} . Suppose $\psi: \mathbb{R} \rightarrow \mathbb{R}$ is convex and $\psi(X)$ is integrable. Then $\psi(E(X | \mathcal{G})) \leq E(\psi(X) | \mathcal{G})$ a.e.

53. Continuation, contraction. If $p \geq 1$, then $\|E(X | \mathcal{G})\|_p \leq \|X\|_p$.

The theory of martingales arose from considerations of gambling systems and has become a staple in the study of stochastic processes. Suppose $\{\mathcal{F}_n, n \geq 0\}$ is a sequence of (sub-) σ -algebras with the property that $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ for each n . Such a sequence is called a *filtration*. A sequence $\{X_n, n \geq 0\}$ of random variables is called a *martingale* with respect to the filtration if, for each n , X_n is integrable, measurable with respect to \mathcal{F}_{n-1} (take $\mathcal{F}_{-1} = \mathcal{F}_0$), and $E(X_n | \mathcal{F}_{n-1}) = X_{n-1}$. The usual filtration is the natural one, $\mathcal{F}_n = \sigma(X_0, X_1, \dots, X_n)$, in which case we write $E(X_n | X_{n-1}) = X_{n-1}$.

54. Basic martingales. Suppose $\{X_n, n \geq 1\}$ is a sequence of independent variables. (a) If the X_n have zero mean, show that the partial sum process formed by setting $S_0 = 0$ and $S_n = X_1 + \dots + X_n$ is a martingale. (b) If the X_n have zero mean and common variance σ^2 then $R_0 = 0$ and $R_n = S_n^2 - n\sigma^2$ is a martingale. (c) If the X_n are positive with unit expectation then $M_0 = 1$ and $M_n = X_1 X_2 \dots X_n$ is a martingale.

55. Pólya's urn. In Pólya's urn scheme of Section II.5, begin with one red and one black ball, sample with replacement and add one ball of the selected colour at each epoch. Let B_n be the number of black balls in the urn after n selection epochs, $M_n = B_n/(n+2)$ the proportion of black balls in the urn. Show that $\{M_n, n \geq 1\}$ is a martingale (relative to a natural filtration that you should specify). Determine the distribution of B_n and thence the distribution of $M = \lim_n M_n$.

56. Martingale transform. If $\{X_n\}$ is a martingale with respect to the filtration $\{\mathcal{F}_n\}$ and, for each n , A_{n+1} is measurable with respect to \mathcal{F}_n , then $Y_0 = X_0$ and $Y_n = X_0 + A_1(X_1 - X_0) + A_2(X_2 - X_1) + \dots + A_n(X_n - X_{n-1})$ is a martingale with respect to the same filtration. A gambler cannot peer into the future to make her bets.

Laplace Transforms

The power of transform methods in analysis was first appreciated by Pierre Simon, Marquis de Laplace in 1785 following earlier work of L. Euler and J. L. Lagrange. Like the Fourier transform that followed it, much of the utility of the Laplace transform may be traced to simple multiplicative properties inherited from the exponential function and the attendant reduction of differential or difference equations to algebraic systems of equations.

1 The transform of a distribution

As Fourier and Laplace transforms rarely occur in the same context, it will be convenient to borrow notation from Section VI.2, both to keep burgeoning notation under control and to emphasise similarities.

DEFINITION The *Laplace transform* of a positive random variable X with distribution F concentrated on the positive half-line $[0, \infty)$ is defined to be the function

$$\widehat{F}(\zeta) = \mathbb{E}(e^{-\zeta X}) = \int_{[0, \infty)} e^{-\zeta x} dF(x) \quad (\zeta \geq 0). \quad (1.1)$$

With a little more attention to precision in terminology, we should say, of course, that \widehat{F} is the Laplace transform of the d.f. F , but the abuse of terminology is standard and harmless. As the integrand is bounded by 1 for all $\zeta \geq 0$ it is clear by monotonicity of expectation that \widehat{F} is well-defined and bounded by 1 for all $\zeta \geq 0$.

The definition takes familiar forms in absolutely continuous or arithmetic cases. If X has density f with support in $[0, \infty)$ then its Laplace transform

$$\widehat{F}(\zeta) = \int_0^\infty f(x)e^{-\zeta x} dx \quad (\zeta \geq 0)$$

is just the ordinary Laplace transform of f that the reader may have seen used as a tool in solving systems of ordinary differential equations.¹ If the positive X is arithmetic with distribution $\{p(k), k \geq 0\}$ then its Laplace transform reduces to the expression

$$\widehat{F}(\zeta) = \sum_{k=0}^{\infty} e^{-\zeta k} p(k) \quad (\zeta \geq 0).$$

In such arithmetic (or discrete) cases it is more usual to set $s = e^{-\zeta}$ and identify $\mathfrak{F}(s) = \widehat{F}(-\log s) = \sum_k s^k p(k)$ with $E(s^X)$, the sum convergent for $0 \leq s \leq 1$. In this form the Laplace transform is called the (*moment*) *generating function* of the distribution $\{p(k)\}$ for reasons that will become apparent shortly. The Laplace transforms of the usual culprits are exhibited in Table 1, the verification requiring nothing more than routine algebraic manoeuvres.

	Distribution $p(k)$ or $f(x)$	Laplace transform $\widehat{F}(\zeta)$
<i>Bernoulli</i>	$p^k q^{1-k}$ ($k = 0, 1$)	$q + pe^{-\zeta}$
<i>Binomial</i>	$\binom{n}{k} p^k q^{n-k}$ ($k \geq 0$)	$(q + pe^{-\zeta})^n$
<i>Poisson</i>	$e^{-\lambda} \lambda^k / k!$ ($k \geq 0$)	$e^{-\lambda + \lambda e^{-\zeta}}$
<i>Geometric</i>	$q^k p$ ($k \geq 0$)	$p(1 - qe^{-\zeta})^{-1}$
<i>Uniform</i>	1 ($0 < x < 1$)	$\zeta^{-1} (1 - e^{-\zeta})$
<i>Exponential</i>	e^{-x} ($x > 0$)	$(1 + \zeta)^{-1}$

Table 1: Laplace transforms of common distributions.

In view of the formal similarity between the Fourier and Laplace transforms it is not at all surprising that the Laplace transform exhibits properties akin to those in Theorem VI.2.1.

Distributions of the same type have Laplace transforms that may be simply related to each other.

THEOREM 1 (CHANGE OF SCALE) *Let $F_a(x) = F(\frac{x}{a})$ be a d.f. of the same type as F obtained by a change of scale $a > 0$. Then $\widehat{F}_a(\zeta) = \widehat{F}(a\zeta)$.*

Verification follows by a simple change of variable inside the Laplace integral. Indeed, if X has d.f. F then $X_a = aX$ has d.f. F_a and accordingly,

$$\widehat{F}_a(\zeta) = E(e^{-\zeta X_a}) = E(e^{-\zeta aX}) = \widehat{F}(a\zeta).$$

¹To obviate confusion some authors call (1.1) the Laplace–Stieltjes transform of F . We will only need the formulation (1.1) and need not make a distinction between the various forms of the transform.

As in the case of the Fourier transform, a key feature of the Laplace transform is the conversion of convolutions into products.

THEOREM 2 (CONVOLUTIONS) *If F_1 and F_2 are d.f.s with support in $[0, \infty)$ then $\widehat{F_1 * F_2}(\zeta) = \widehat{F_1}(\zeta)\widehat{F_2}(\zeta)$. A fortiori, for any d.f. F with support in the positive half-line $[0, \infty)$, we have $\widehat{F^{*n}}(\zeta) = \widehat{F}(\zeta)^n$ for each positive integer n .*

Again, matters are simplest in the expectation formulation. Suppose X_1 and X_2 are independent, positive random variables with d.f.s F_1 and F_2 , respectively. Then the variable $S = X_1 + X_2$ has d.f. $F_1 * F_2$ and

$$\widehat{F_1 * F_2}(\zeta) = \mathbf{E}(e^{-\zeta S}) = \mathbf{E}(e^{-\zeta X_1} e^{-\zeta X_2}) \stackrel{(*)}{=} \mathbf{E}(e^{-\zeta X_1}) \mathbf{E}(e^{-\zeta X_2}) = \widehat{F_1}(\zeta)\widehat{F_2}(\zeta),$$

the key step (*) following by the independence of the random variables $e^{-\zeta X_1}$ and $e^{-\zeta X_2}$ as independent variables are uncorrelated.

EXAMPLES: 1) *Again, the binomial distribution.* If $F(x) = qH_0(x) + pH_0(x-1)$ is the Bernoulli distribution with success probability p then $F^{*n}(x)$ is the binomial distribution corresponding to the accumulated successes in n tosses of a coin with success probability p . We hence obtain $\widehat{F^{*n}}(\zeta) = \widehat{F}(\zeta)^n = (q + pe^{-\zeta})^n$ matching the direct calculation reported in Table 1.

2) *The negative binomial distribution.* The waiting time for the r th success in repeated tosses of a coin with success probability p has the negative binomial distribution $\binom{r}{k}(-q)^k p^r$ which may be expressed as the r -fold convolution of the geometric distribution $q^k p$ with itself. Accordingly, the Laplace transform of the negative binomial distribution is given by $p^r(1 - qe^{-\zeta})^{-r}$.

3) *Convolutions of the uniform distribution.* Let $U^{*n}(x)$ be the n -fold convolution of the uniform distribution $U(x)$ on the unit interval. Then

$$\widehat{U^{*n}}(\zeta) = \left(\frac{1 - e^{-\zeta}}{\zeta} \right)^n = \sum_{k=0}^n (-1)^k \binom{n}{k} e^{-k\zeta} \zeta^{-n},$$

the final step courtesy the binomial expansion. This expression may be used as a starting point for a rederivation of the expression (IX.1.2) of the associated density u_n but we will not pursue it here.

4) *The gamma density.* The gamma distribution $G_n(x) = 1 - e^{-x} \sum_{k=0}^{n-1} x^k / k!$ ($x > 0$) is obtained via the n -fold convolution of the exponential distribution with mean 1. The associated Laplace transform is hence $\widehat{G_n}(\zeta) = (1 + \zeta)^{-n}$. ►

Repeated formal differentiations of both sides of (1.1) result in very useful identities.

THEOREM 3 (DERIVATIVES) Suppose F is a d.f. with support on the positive half-line $[0, \infty)$, \widehat{F} the associated Laplace transform. Then $\widehat{F}(\zeta)$ has derivatives of all orders for $\zeta > 0$ and

$$\frac{d^n}{d\zeta^n} \widehat{F}(\zeta) = \widehat{F}^{(n)}(\zeta) = (-1)^n \int_{[0, \infty)} x^n e^{-\zeta x} dF(x) \quad (1.2)$$

for each $n \geq 0$.

The proof requires little more than the fact that the function $x^n e^{-\zeta x}$ is twice continuously differentiable for each fixed $\zeta > 0$ and each n to justify differentiating under the integral sign.

For the reader who would like to embellish the bald sentence above, the case $n = 0$ is just the definition of the Laplace transform and establishes the base of an induction. As induction hypothesis, suppose now that (1.2) holds for some n . Then, by additivity of expectation,

$$\begin{aligned} \frac{\widehat{F}^{(n)}(\zeta + h) - \widehat{F}^{(n)}(\zeta)}{h} - \int_{[0, \infty)} (-x)^{n+1} e^{-\zeta x} dF(x) \\ = \int_{[0, \infty)} (-x)^n e^{-\zeta x} \left(\frac{e^{-hx} - 1 + hx}{h} \right) dF(x). \end{aligned}$$

By Taylor's theorem (or repeated applications of the mean value theorem), if f is twice continuously differentiable then $f(x) = f(0) + f'(0)x + f''(\xi)x^2/2$ for some $\xi = \xi(x)$ lying between 0 and x . Identifying $f(x) = e^{-hx}$, it follows that $e^{-hx} - 1 + hx = e^{-h\xi} h^2 x^2/2$ for some $\xi = \xi(x)$, whence $|e^{-hx} - 1 + hx|/h \leq hx^2/2$. It follows that

$$\begin{aligned} & \left| \frac{\widehat{F}^{(n)}(\zeta + h) - \widehat{F}^{(n)}(\zeta)}{h} - \int_{[0, \infty)} (-x)^{n+1} e^{-\zeta x} dF(x) \right| \\ & \leq \int_{[0, \infty)} x^n e^{-\zeta x} \left| \frac{e^{-hx} - 1 + hx}{h} \right| dF(x) \leq \frac{h}{2} \int_{[0, \infty)} x^{n+2} e^{-\zeta x} dF(x). \end{aligned}$$

The integral on the right converges for all $\zeta > 0$ in view of the rapid extinction of the exponential in the tails. Allowing h to tend to zero then verifies the form of the $(n+1)$ th derivative of \widehat{F} and completes the induction. ►

If we allow ζ to tend to zero from the right in (1.2) we obtain formal expressions for the moments of the associated random variable when they exist: if X has an n th moment $\mu_n = E(X^n)$ then $\mu_n = (-1)^n \widehat{F}^{(n)}(0+)$ and a fortiori $\mu = \mu_1 = -\widehat{F}'(0+)$. The Laplace transform of the distribution hence determines all its moments. The converse is also true provided the sequence $\{\mu_n\}$ does not have too rapid a rate of growth. The following simple sufficient condition for the validity of the converse statement, for instance, is an easy consequence of Taylor's theorem.

Suppose X is positive and has all moments μ_n . By Taylor's expansion of \widehat{F} through n terms we have

$$\widehat{F}(\zeta) = \sum_{k=0}^{n-1} \widehat{F}^{(k)}(0) \frac{\zeta^k}{k!} + \frac{1}{(n-1)!} \int_0^\zeta \xi^{n-1} \widehat{F}^{(n)}(\zeta - \xi) d\xi.$$

(If the reader does not recall this she can verify it quickly by induction by a repeated integration by parts of the integral on the right.) By (1.2), we have $\widehat{F}^{(k)}(0) = (-1)^k \mu_k$ for each k and $0 \leq |\widehat{F}^{(n)}(\zeta)| \leq \mu_n$, and so we obtain

$$\left| \widehat{F}(\zeta) - \sum_{k=0}^{n-1} \frac{(-1)^k \mu_k \zeta^k}{k!} \right| \leq \frac{\mu_n}{(n-1)!} \int_0^\zeta \xi^{n-1} d\xi = \frac{\mu_n \zeta^n}{n!}.$$

We could use Stirling's formula to estimate the factorial in the denominator on the right but the simple bound $n! \geq n^n e^{-n}$ [see (IV.5.2)] suffices. For the bound on the right to tend to zero asymptotically it hence suffices if $\frac{1}{n} \mu_n^{1/n} \rightarrow 0$.

THEOREM 4 *If $\frac{1}{n} \mu_n^{1/n} \rightarrow 0$ then the series $\sum_{k=0}^{\infty} (-1)^k \mu_k \zeta^k / k!$ converges pointwise to $\widehat{F}(\zeta)$.*

One of the reasons for the utility of the Laplace transform \widehat{F} is that it completely determines the underlying d.f. F .

AN INVERSION THEOREM *At each point of continuity of F we have*

$$\sum_{n \leq \zeta x} \frac{(-\zeta)^n}{n!} \widehat{F}^{(n)}(\zeta) \rightarrow F(x) \quad (\zeta \rightarrow \infty). \quad (1.3)$$

In particular, F is uniquely determined by its Laplace transform.

A feeling for the nature of the result can be obtained by leveraging (1.2) and linearity of expectation to write

$$\sum_{n \leq \zeta x} \frac{(-\zeta)^n}{n!} \widehat{F}^{(n)}(\zeta) = \int_{[0, \infty)} \sum_{n \leq \zeta x} \frac{(\zeta u)^n}{n!} e^{-\zeta u} dF(u).$$

Identifying the summands $p(n; \zeta u) = \frac{(\zeta u)^n}{n!} e^{-\zeta u}$ on the right with the Poisson distribution with mean ζu , it becomes clear that the nature of the convergence is determined by the asymptotic behaviour of the Poisson distribution. As, for each u , the distribution $p(n; \zeta u)$ has variance ζu , for large ζ it is effectively concentrated around its mean value of ζu . Accordingly, the sum inside the integral is close to either 1 or 0 depending on whether $u < x$ or $u > x$.

LEMMA 1 For every $x > 0$, asymptotically as $\zeta \rightarrow \infty$, we have

$$\sum_{n \leq \zeta x} p(n; \zeta u) \rightarrow \begin{cases} 1 & \text{if } u < x, \\ 0 & \text{if } u > x. \end{cases}$$

PROOF: Let X be a Poisson random variable with mean λ . Writing $p(n; \lambda) = e^{-\lambda} \lambda^n / n!$ ($n \geq 0$) for the distribution of X as usual, for any $0 < \delta < 1$, we have

$$\begin{aligned} P\{|X - \lambda| \geq \delta \lambda\} &= \sum_{|n - \lambda| \geq \delta \lambda} p(n; \lambda) \leq \sum_{|n - \lambda| \geq \delta \lambda} \left(\frac{n - \lambda}{\delta \lambda} \right)^2 p(n; \lambda) \\ &\leq \frac{1}{\delta^2 \lambda^2} \sum_{n=0}^{\infty} (n - \lambda)^2 p(n; \lambda) = \frac{\text{Var}(X)}{\delta^2 \lambda^2} = \frac{1}{\delta^2 \lambda}. \end{aligned}$$

The reader may recognise the Chebyshev trick in wresting the sum into a manageable upper bound. Identifying λ with ζu completes the proof. ►

Our lemma suggests a reduction of the problem to a comparison of the Poisson d.f. $P_\zeta(u, x) = \sum_{n \leq \zeta x} p(n; \zeta u)$ and the shifted and coordinate-reversed Heaviside function $Q(u, x) = H_0(-(u - x))$ which, viewing x as a parameter, takes value 1 when $u \leq x$ and value 0 otherwise. The following monotonicity result for the Poisson distribution $p(n; \lambda) = e^{-\lambda} \lambda^n / n!$ ($n \geq 0$) which the reader may well feel is natural paves the way.

LEMMA 2 If $0 < \lambda_1 < \lambda_2$ then $\sum_{n=0}^m p(n; \lambda_1) \geq \sum_{n=0}^m p(n; \lambda_2)$ for every m .

PROOF: A direct algebraic assault is painful but the result becomes transparent if the reader keeps the probabilistic setting plainly in mind. Let X_1 and X be independent Poisson variables of means λ_1 and $\lambda = \lambda_2 - \lambda_1$, respectively. Then $X_2 = X_1 + X$ is a Poisson variable with mean $\lambda_1 + \lambda = \lambda_2$ as a consequence of the closure of the Poisson distribution under convolution. As X is manifestly positive, we have $X_1 \leq X_2$, whence $P\{X_1 \leq m\} \geq P\{X_2 \leq m\}$ for each m . ►

In jargon we say that X_1 is *stochastically dominated* by X_2 , the terminology adding colour to the bald inequality.

PROOF OF THE INVERSION THEOREM: Suppose $x > 0$ is a point of continuity of the d.f. F . Then to each fixed $\epsilon > 0$ we may select $0 < \delta < x$ sufficiently small so that $F(x + \delta) - F(x - \delta) < \epsilon$. To obviate notational nuisances we may as well suppose that $x - \delta$ and $x + \delta$ are both points of continuity of F . A comparison of the two sides of (1.3) now yields

$$\left| \sum_{n \leq \zeta x} \frac{(-\zeta)^n}{n!} \hat{F}^{(n)}(\zeta) - F(x) \right| = \left| \int_{[0, \infty)} P_\zeta(u, x) dF(u) - \int_{[0, \infty)} Q(u, x) dF(u) \right|$$

$$\leq \int_{|u-x|<\delta} |P_\zeta(u, x) - Q(u, x)| dF(u) + \int_{|u-x|\geq\delta} |P_\zeta(u, x) - Q(u, x)| dF(u).$$

As $P_\zeta(u, x)$ and $Q(u, x)$ are both bounded above by 1 the first integral on the right is bounded above by $2 \int_{x-\delta}^{x+\delta} dF(u) = 2[F(x+\delta) - F(x-\delta)] < 2\epsilon$. For the second integral, by virtue of Lemma 1 and Lemma 2,

$$\begin{aligned} \max\{|P_\zeta(u, x) - Q(u, x)| : u \leq x - \delta \text{ or } u \geq x + \delta\} \\ \leq \max\{1 - P_\zeta(x - \delta, x), P_\zeta(x + \delta, x) - 0\} < \epsilon \end{aligned}$$

for sufficiently large ζ . It follows that the contribution for values u outside the immediate neighbourhood of x is bounded above by $\epsilon \int_{|u-x|\geq\delta} dF(u) \leq \epsilon \int_{[0, \infty)} dF(u) = \epsilon$. Accordingly,

$$\left| \sum_{n \leq \zeta x} \frac{(-\zeta)^n}{n!} \widehat{F}^{(n)}(\zeta) - F(x) \right| < 3\epsilon$$

for all sufficiently large ζ and as $\epsilon > 0$ may be chosen arbitrarily small this completes the proof of the theorem. ▶

Coupled with Theorem 4, our inversion theorem provides cover for the following

SLOGAN *The moments μ_n of X determine its distribution F if the sequence $\{\mu_n\}$ does not grow too rabidly with n .*

The simple condition of Theorem 4 suffices for our purposes in this volume though it is not the strongest result of its type. The best result is due to Carleman: *the distribution F is determined by its moments if the series $\sum_n \mu_n^{-1/n}$ diverges.*

Inversion is simple when the underlying distribution is discrete. Suppose $G(s) = E(s^X)$ is the generating function of a positive, arithmetic random variable X . The reader is no doubt aware that power series expansions are unique and, accordingly, we may identify the underlying distribution $\{p(k), k \geq 0\}$ uniquely by expanding G in a power series, $G(s) = \sum_{k=0}^{\infty} p(k)s^k$, and comparing terms.

2 Extensions

The main properties of the Laplace transform of distributions concentrated on the positive half-line carry through without fuss to the Laplace transform of measures engendered by increasing, right-continuous functions concentrated on the positive half-line. Suppose $V(x)$ is a monotonically increasing, right-continuous function, not necessarily bounded, and vanishing for $x < 0$. In

notation, $V(x) \leq V(y)$ if $x < y$ and $V(x) = 0$ if $x < 0$. In the usual abuse of notation, let $V(\mathbb{I})$ represent the measure induced by $V(x)$ on the Borel sets of the line; as usual, we refer to $V(x)$ as the *distribution* associated with this measure though it need not be a probability distribution. The measure $V(\mathbb{I})$ is concentrated on the positive half-line $\mathbb{R}^+ = [0, \infty)$ and while it may not be finite it does attach finite measure $V(a, b] = V(b) - V(a)$ to every finite interval $(a, b]$. An integral of the form

$$\widehat{V}(\zeta) = \int_{\mathbb{R}^+} e^{-\zeta x} V(dx) \quad (2.1)$$

now constitutes a natural generalisation of (1.1) to measures that are not necessarily finite but we now run into the irritating possibility that the integral on the right may not converge. If it does converge for some $\zeta = \zeta_0$, then it converges for all $\zeta > \zeta_0$ by monotonicity.

DEFINITION If the integral in (2.1) converges for some $\zeta = \zeta_0$, then the function $\widehat{V}(\zeta)$ defined for $\zeta > \zeta_0$ is called the *Laplace transform* of the measure V .

The properties of the Laplace transform that we have seen hitherto continue to hold in the general setting with the caveat that the transform is now defined on some subinterval of the positive half-line.

THEOREM 1' If $\widehat{V}(\zeta)$ is defined for $\zeta > \zeta_0$ and $V_a(x) = V(\frac{x}{a})$ is obtained via a change of scale for any $a > 0$, then $\widehat{V}_a(\zeta) = \widehat{V}(a\zeta)$ is convergent for $\zeta > \zeta_0/a$.

THEOREM 2' Suppose the distributions U and V have Laplace transforms convergent for $\zeta > \zeta_0$. Then so does the Laplace transform of the convolution $U \star V$ and, moreover, $\widehat{U \star V}(\zeta) = \widehat{U}(\zeta)\widehat{V}(\zeta)$.

THEOREM 3' If the distribution V has a Laplace transform convergent on an interval then its Laplace transform \widehat{V} has derivatives of all orders on that interval and, moreover,

$$\widehat{V}^{(n)}(\zeta) = (-1)^n \int_{\mathbb{R}^+} x^n e^{-\zeta x} V(dx).$$

A GENERAL INVERSION THEOREM Suppose V is a distribution, not necessarily bounded, concentrated on the positive half-line and suppose its Laplace transform \widehat{V} exists in some interval $[\zeta_0, \infty)$. Then V is determined uniquely by \widehat{V} .

The proofs are largely unchanged with the replacement of stray references to probability distributions by general distributions with convergent Laplace transforms. The change of scale property follows by a change of variable inside the Laplace integral. For the convolution property we exploit the natural generalisation of the idea of independent functions: if the pair (u, v) is

equipped with product measure $U \otimes V$ on the Euclidean plane then $w = u + v$ has measure $W = U \star V$ and the result follows by Fubini's theorem,

$$\begin{aligned} \int_{\mathbb{R}^+} e^{-\zeta w} W(dw) &= \iint_{\mathbb{R}^+ \times \mathbb{R}^+} e^{-\zeta(u+v)} U \otimes V(du, dv) \\ &= \int_{\mathbb{R}^+} e^{-\zeta u} U(du) \int_{\mathbb{R}^+} e^{-\zeta v} V(dv). \end{aligned}$$

To verify the derivative property it suffices to observe that, if $\int_{\mathbb{R}^+} e^{-\zeta x} V(dx)$ is convergent for $\zeta > a$, then so is $\int_{\mathbb{R}^+} x^n e^{-\zeta x} V(dx)$ for every positive integer n . Indeed, for any $n \geq 0$ and any $a < \zeta' < \zeta$, we may select a positive constant A such that $x^n e^{-\zeta x} \leq A e^{-\zeta' x}$ for all $x \geq 0$. It follows by monotonicity of integration that $\int_{\mathbb{R}^+} x^n e^{-\zeta x} V(dx) \leq A \int_{\mathbb{R}^+} e^{-\zeta' x} V(dx)$ is convergent. The rest of the proof remains unaltered.

The inversion property is the only one that requires a little work. Suppose $\widehat{V}(\zeta_0)$ exists for some ζ_0 . Then the normalised measure V^* defined on each Borel set \mathbb{I} by

$$V^*(\mathbb{I}) = \int_{\mathbb{I}} \frac{e^{-\zeta_0 x}}{\widehat{V}(\zeta_0)} V(dx), \quad \text{or, in short, } V^*(dx) = \frac{e^{-\zeta_0 x}}{\widehat{V}(\zeta_0)} V(dx)$$

is bounded and indeed is a probability measure as $V^*(\mathbb{R}) = 1$. The corresponding Laplace transform $\widehat{V}^*(\zeta)$ is hence well-defined for all $\zeta \geq 0$ and, by the basic uniqueness theorem for the Laplace transform of probability distributions, uniquely determines the underlying probability distribution V^* . But the monotonicity of the defining relation between distributions implies that V^* uniquely determines V , indeed, $V(dx) = \widehat{V}(\zeta_0) e^{\zeta_0 x} V^*(dx)$, so that it follows that the distribution V is uniquely determined by the Laplace transform \widehat{V}^* of the normalised measure. On the other hand,

$$\widehat{V}^*(\zeta) = \int_{\mathbb{R}^+} e^{-\zeta x} V^*(dx) = \int_{\mathbb{R}^+} e^{-\zeta x} \frac{e^{-\zeta_0 x}}{\widehat{V}(\zeta_0)} V(dx) = \frac{\widehat{V}(\zeta_0 + \zeta)}{\widehat{V}(\zeta_0)} \quad (\zeta \geq 0)$$

is completely determined by \widehat{V} . It follows, as asserted, that the distribution V is completely determined by its Laplace transform \widehat{V} .

A final feature worth remarking is that linearity of integration shows that the Laplace transform is additive.

THEOREM 5' (LINEARITY) *Suppose V_1 and V_2 are distributions concentrated on the positive half-line with Laplace transforms convergent on some interval (ζ_0, ∞) . Let V be the mixture distribution $V = \lambda_1 V_1 + \lambda_2 V_2$ for any positive λ_1 and λ_2 . Then the associated measure has a Laplace transform $\widehat{V}(\zeta) = \lambda_1 \widehat{V}_1(\zeta) + \lambda_2 \widehat{V}_2(\zeta)$ converging for $\zeta > \zeta_0$.*

I have included the proof though the reader should by now be able to walk through the steps of the Lebesgue programme on her own. We can in fact show a little more than is claimed. Suppose g is integrable with respect to V_1 and V_2 . Then $\int g d(\lambda_1 V_1 + \lambda_2 V_2) = \lambda_1 \int g dV_1 + \lambda_2 \int g dV_2$. The proof of the assertion proceeds in the usual sequence. If $g = 1_{\mathbb{A}}$ is the indicator of a Borel set \mathbb{A} then the assertion is equivalent to the defining equation $V(\mathbb{A}) = \lambda_1 V_1(\mathbb{A}) + \lambda_2 V_2(\mathbb{A})$. We next proceed to a simple function $g = \sum_{k=1}^m c_k 1_{\mathbb{A}_k}$ whence the claimed result holds by additivity of integration. The result hence holds for each simple g_n increasing to the positive g . Three applications of the monotone convergence theorem then show that the result holds for positive g as well. We complete the proof by decomposing any general measurable function $g = g^+ - g^-$ as a difference of its positive and negative parts. Additivity of the Laplace transform follows by identifying $g(x) = e^{-\zeta x}$.

3 The renewal equation and process

Suppose F is a bounded distribution concentrated on the positive half-line (or, equivalently, λF is a probability distribution for some positive λ) and G is a distribution, not necessarily bounded, also concentrated on the positive half-line. We then say that a distribution $V(t)$, vanishing for $t < 0$, is a solution of the *renewal equation* if it satisfies the convolutional relationship

$$V(t) = G(t) + (F * V)(t), \text{ or, spelled out, } V(t) = G(t) + \int_{[0,t]} V(t-x) dF(x). \quad (3.1)$$

If the Laplace transform of G exists, say, in the interval (ζ_0, ∞) , and F is not concentrated at the origin, then, by formally taking Laplace transforms of both sides we obtain

$$\widehat{V}(\zeta) = \widehat{G}(\zeta) + \widehat{V}(\zeta)\widehat{F}(\zeta), \text{ or, equivalently, } \widehat{V}(\zeta) = \frac{\widehat{G}(\zeta)}{1 - \widehat{F}(\zeta)} = \widehat{G}(\zeta) \sum_{k=0}^{\infty} \widehat{F}(\zeta)^k, \quad (3.2)$$

for all $\zeta > \zeta_0$. As \widehat{G} and \widehat{F} have derivatives of all orders, it follows that so does \widehat{V} , but more can be said. In view of the derivative property of the Laplace transform, it is clear that $(-1)^n \widehat{F}(\zeta)$ and $(-1)^n \widehat{G}(\zeta)$ are both positive for each n . Functions with this character are extremely regular as discovered by S. Bernstein.

DEFINITION A function ψ on an interval (ζ_0, ∞) is *completely monotone* if it possesses derivatives $\psi^{(n)}$ of all orders $n \geq 0$ and $(-1)^n \psi^{(n)}(\zeta) \geq 0$ for all $\zeta > \zeta_0$.²

²As a notational convention, we set $\psi^{(0)}(\zeta) = \psi(\zeta)$.

Thus, a completely monotone function is positive and has derivatives of alternating sign. In this terminology, the derivative property of the Laplace transform asserts that the Laplace transform of a distribution is completely monotone. Remarkably, as discovered by Bernstein, the converse is also true: *a function is completely monotone if, and only if, it is the Laplace transform of a distribution.* I shall defer the proof of this beautiful result to Section XIX.7 as it is best seen through the prism of Helly's selection principle.

Two preliminary results will now clear the way for a return to the renewal equation.

LEMMA 1 Suppose ψ is completely monotone and $\psi(\zeta) < 1$ for all $\zeta > \zeta_0$. Then $u(\zeta) = (1 - \psi(\zeta))^{-1}$ is completely monotone on (ζ_0, ∞) .

PROOF: Set $p(\zeta) = 1 - \psi(\zeta)$. Then $p(\zeta) = p^{(0)}(\zeta) > 0$ and, for $n \geq 1$, $p^{(n)}(\zeta) = -\psi^{(n)}(\zeta)$. It is now simplest to clear the denominator in $u(\zeta) = 1/p(\zeta)$ to avoid a painful differentiation of fractions. Beginning with the identity $p(\zeta)u(\zeta) = 1$, by repeated differentiation of both sides we obtain

$$\sum_{k=0}^n \binom{n}{k} p^{(n-k)}(\zeta) u^{(k)}(\zeta) = 0, \text{ or, } u^{(n)}(\zeta) = \frac{1}{p(\zeta)} \sum_{k=0}^{n-1} \binom{n}{k} \psi^{(n-k)}(\zeta) u^{(k)}(\zeta).$$

The stage is set for an induction. The induction base is provided by the obvious inequality $u^{(0)}(\zeta) = u(\zeta) = 1/p(\zeta) > 0$, and we now obtain

$$(-1)^n u^{(n)}(\zeta) = \frac{1}{p(\zeta)} \sum_{k=0}^{n-1} \binom{n}{k} ((-1)^{n-k} \psi^{(n-k)}(\zeta)) ((-1)^k u^{(k)}(\zeta)) > 0$$

by induction hypothesis. ▶

LEMMA 2 The product of completely monotone functions is also completely monotone.

PROOF: Suppose φ and ψ are completely monotone (on some interval) and $\vartheta = \varphi\psi$. By repeated differentiation, $\vartheta^{(n)} = \sum_{k=0}^n \binom{n}{k} \varphi^{(n-k)} \psi^{(k)}$, whence

$$(-1)^n \vartheta^{(n)}(\zeta) = \sum_{k=0}^n \binom{n}{k} ((-1)^{n-k} \varphi^{(n-k)}(\zeta)) ((-1)^k \psi^{(k)}(\zeta)).$$

The claimed result follows as each of the summands is positive. ▶

In view of (3.2), it follows that \widehat{V} is completely monotone, hence by Bernstein's theorem the Laplace transform of some distribution V . But then the general inversion theorem of Section 2 shows that V is uniquely determined by \widehat{V} . It follows that *the renewal equation (3.1) has a unique solution.*

The prototypical occurrence of the renewal equation is in the renewal process. Suppose X_1, X_2, \dots are independent, positive random variables with a common d.f. $F(x)$. To obviate notational nuisances, we suppose that $F(0) = 0$ so that F puts no mass at the origin. We think of each X_j as the time between *renewals* of some phenomenon. Typical examples include arrivals of customers into a queue, the duration between epochs when a queue empties out, and returns of a random walk to the origin. Let $S_n = X_1 + \dots + X_n$ denote the *waiting time till the nth renewal*, the terminology being self-explanatory. Then S_n is positive with d.f. $F^{*n}(x)$.

If we think of a process unfolding in time with intervals punctuated by renewals (or arrivals), then the *renewal epochs* S_n connote points in time when the process probabilistically restarts. A variety of useful features can be deduced from this viewpoint.

The sequence of waiting times $\{S_n, n \geq 1\}$ determines the basic *renewal process* $N(t) = \sup\{n : S_n \leq t\}$ which connotes the number of renewals in the interval $(0, t]$. Thus, if $\{X_j, j \geq 1\}$ represents the time between arrivals of customers to a queue, then $N(t)$ represents the number of customers who have arrived up till time t .

By convention we suppose that the process begins with a renewal at time zero, that is to say, $N(0) = 1$, and set $N(t) = 0$ for $t < 0$. It will be convenient now to extend notation to write $N(\mathbb{I}) = \sum_n 1(S_n \in \mathbb{I})$ for the number of renewal epochs in any given interval \mathbb{I} ; naturally enough, we set $N(\emptyset) = 0$. By focusing on the time of the first renewal epoch, the number of renewals in the interval $[0, t]$ is given by $N[0, t] = 1 + N[X_1, t]$. Now, by the independence of the inter-renewal times, if x is a renewal epoch, then the number of renewals in the interval $[x, t]$ is a probabilistic replica of the number of renewals in any interval of duration $t - x$ beginning with a renewal. By integrating with respect to the first renewal epoch, the expected number of renewals up till time $t > 0$ is accordingly given by

$$U(t) = E(N[0, t]) = 1 + E(N[X_1, t]) = 1 + E(N[0, t - X_1]) = 1 + \int_{[0, t]} U(t - x) dF(x).$$

As $U(t - x) = 0$ for $x > t$ and $F(x) = 0$ for $x \leq 0$, we may extend the domain of integration on the right to the entire real line without changing the integral value. In other words,

$$U(t) = H_0(t) + \int_{\mathbb{R}} U(t - x) dF(x) = H_0(t) + (F * U)(t) \quad (t > 0) \quad (3.3)$$

satisfies the renewal equation with $G(t) = H_0(t)$ the Heaviside distribution representing a point mass concentrated at the origin. As $\widehat{H}_0(\zeta) = 1$, the second equation in (3.2) specialises to the simple form

$$\widehat{U}(\zeta) = [1 - \widehat{F}(\zeta)]^{-1} \quad (\zeta > 0). \quad (3.3')$$

The expected number of renewals $U(t)$ is called the *renewal function*. It is clear that $U(t)$ is increasing, and hence determines a distribution on the real line with $U(\mathbb{I})$ representing the expected number of renewals in the Borel set \mathbb{I} .

In the discrete case it is preferable to deal directly with point probabilities and the convolution integral segues into a sum. Suppose F is arithmetic and places mass $f(k)$ at integer values $k > 0$. (For definiteness, we set $f(k) = 0$ for $k \leq 0$.) Then renewals can occur only at positive integer epochs and, for integer $n \geq 0$, (3.3) becomes

$$U(n) = 1 + \sum_{k=0}^{\infty} U(n-k)f(k).$$

Writing $u(n)$ for the probability that there is a renewal at epoch $n \geq 0$, we have $U(n) = u(0) + u(1) + \dots + u(n)$ with the natural conventions $u(0) = 1$ and $u(n) = 0$ for $n < 0$. The specialisation of (3.3) to point probabilities results in a convolution sum

$$\begin{aligned} u(n) = U(n) - U(n-1) &= \sum_{k=0}^{\infty} [U(n-k) - U(n-1-k)]f(k) \\ &= \sum_{k=0}^{\infty} u(n-k)f(k) = (f * u)(n) \quad (n \geq 1) \end{aligned} \quad (3.4)$$

for the renewal probabilities. The sum on the right is only formally infinite with non-zero contributions only for $1 \leq k \leq n$ as $f(0) = 0$ and $u(n-k) = 0$ for $k > n$. If $\mathfrak{U}(s) = \sum_{n=0}^{\infty} u(n)s^n$ and $\mathfrak{F}(s) = \sum_{n=0}^{\infty} f(n)s^n$ represent the generating functions of the sequences $\{u(n)\}$ and $\{f(n)\}$, respectively, then, with $s = e^{-\zeta}$, the relation (3.3') becomes

$$\mathfrak{U}(s) = [1 - \mathfrak{F}(s)]^{-1} \quad (0 \leq s < 1). \quad (3.4')$$

The reader who is following along critically will have realised that in order to be able to proceed from (3.3) to (3.3') it is requisite that we show that $U(t)$ is finite for every t . We now show that this is the case. As the event $N(t) \geq n$ occurs if, and only if, $S_n \leq t$, we obtain courtesy Theorem XIII.5.4 that

$$U(t) = 1 + E(N(t)) = 1 + \sum_{n=1}^{\infty} P\{N(t) \geq n\} = 1 + \sum_{n=0}^{\infty} P\{S_n \leq t\} = \sum_{n=0}^{\infty} F^{*n}(t), \quad (3.5)$$

where the convention $F^{*0}(t) = 1$ helps compact expressions. By the monotonicity of F ,

$$F^{*n}(t) = \int_{[0,t]} F(t-x) dF^{*(n-1)}(x) \leq F(t) \int_{[0,t]} dF^{*(n-1)}(x) = F(t)F^{*(n-1)}(t),$$

so that by induction it follows that $F^{*n}(t) \leq F(t)^n$ for all $n \geq 0$. If $F(t) < 1$ for all t then the series on the right of (3.5) is dominated by a geometric series with term $F(t) < 1$

and accordingly is finite. The argument does not hold as given if F is concentrated in a finite interval as $F(t)$ may be identically one. But, in this case, there exists $t_0 > 0$ with $F(t_0) < 1$. Then

$$(F * F)(2t_0) = \int_{[0, 2t_0]} F(2t_0 - x) dF(x) \leq \int_{[0, t_0]} dF(x) + F(t_0) \int_{(t_0, 2t_0]} dF(x)$$

as $F(2t_0 - x) \leq 1$ for all x and certainly also in the interval $[0, t_0]$, while in the interval $t_0 < x \leq 2t_0$, $F(2t_0 - x) \leq F(t_0)$ by monotonicity of the d.f. It follows that $(F * F)(2t_0) \leq F(t_0) + F(t_0)[F(2t_0) - F(t_0)] \leq F(t_0)[2 - F(t_0)]$. The function $g(x) = x(2 - x)$ achieves its unique maximum value of 1 at $x = 1$ as is easily verified by differentiation. Identifying x with $F(t_0)$ it follows that $(F * F)(2t_0) \leq g(F(t_0)) < 1$ as $F(t_0) < 1$. By induction we hence obtain that $F^{*2^k}(2^k t_0) < 1$ for each $k \geq 0$. The reader may prefer the following slogan which captures the essence of what we have discovered.

SLOGAN *Convolution has a dispersive or smoothing character.*

For any given $t > 0$ we may now select $m = 2^k$ sufficiently large so that $mt_0 > t$, whence $F^{*m}(t) \leq F^{*m}(mt_0) < 1$. As $F^{*j}(t) \geq F^{*k}(t)$ if $j \leq k$, we may now bound the right-hand side of (3.5) by considering blocks of m renewals at a time to obtain

$$U(t) \leq m \sum_{k=0}^{\infty} F^{*m^k}(t) \leq m \sum_{k=0}^{\infty} F^{*m}(t)^k \leq m \sum_{k=0}^{\infty} F^{*m}(mt_0)^k,$$

the geometric series on the right again convergent as $F^{*m}(mt_0) < 1$. It follows, as asserted, that $U(t)$ is finite for all t .

The renewal equation crops up in a bewildering variety of situations. The reader may get a feeling for its ubiquity—and the reason for its name—in the applications that follow.

4 Gaps in the Poisson process

The most important renewal process arises when the inter-renewal times are governed by an exponential distribution. By virtue of the memoryless property of the exponential distribution, this is a natural model to capture randomness in the arrival of customers to a queue (or packets to an internet router or cars to an intersection). Suppose, for definiteness, that the inter-renewal times X_1, X_2, \dots are independent and governed by the standard exponential distribution of mean one. The induced renewal process $N(t)$ is then the Poisson process of Section IX.7. This is the prototypical renewal process.

Let W_τ denote the waiting time till the first gap of size (at least) τ is observed between successive renewals and let $V(t) = P\{W_\tau \leq t\}$ be its d.f. Clearly, $W_\tau \geq \tau$, whence $V(t) = 0$ for $t < \tau$. For $t \geq \tau$, the event $W_\tau \leq t$ occurs if either the first renewal occurs after time τ (an event of probability $e^{-\tau}$), or the first renewal occurs before τ , say at x , and there is a gap of size $\geq \tau$ between x

and t in which there are no renewals. Integrating over the possible values of x we obtain, for each $t > \tau$,

$$V(t) = e^{-\tau} H_0(t - \tau) + \int_0^\tau V(t - x) e^{-x} dx = G(t) + (F * V)(t),$$

where $G(t) = e^{-\tau} H_0(t - \tau)$ represents the point measure of mass $e^{-\tau}$ concentrated at $t = \tau$ and $F(x)$ is the truncated exponential measure concentrated in the interval $0 < x < \tau$ with density given by $e^{-x} 1_{(0,\tau)}(x)$. The upper limit of the convolution integral is dictated by the fact that the density of F is identically zero for $x > \tau$. Taking Laplace transforms, we have

$$\widehat{G}(\zeta) = \int_{\mathbb{R}^+} e^{-\zeta t} dG(t) = \int_{\mathbb{R}^+} e^{-\zeta t} e^{-\tau} dH_0(t - \tau) = e^{-(1+\zeta)\tau},$$

as integration with respect to $H_0(t - \tau)$ merely picks out the value of the integrand at $t = \tau$, while a simple exponential integral shows that

$$\widehat{F}(\zeta) = \int_{\mathbb{R}^+} e^{-\zeta x} dF(x) = \int_0^\tau e^{-\zeta x} e^{-x} dx = \frac{1 - e^{-(1+\zeta)\tau}}{1 + \zeta}.$$

It follows by (3.2) that

$$\widehat{V}(\zeta) = \frac{\widehat{G}(\zeta)}{1 - \widehat{F}(\zeta)} = \frac{(1 + \zeta)e^{-(1+\zeta)\tau}}{\zeta + e^{-(1+\zeta)\tau}} \quad (\zeta > 0).$$

The moments of W_τ can now be obtained by differentiation. Clearing the denominator to avoid a tedious differentiation of fractions, we obtain

$$\begin{aligned} \widehat{V}(\zeta)(\zeta + e^{-(1+\zeta)\tau}) &= (1 + \zeta)e^{-(1+\zeta)\tau}, \\ \widehat{V}'(\zeta)(\zeta + e^{-(1+\zeta)\tau}) + \widehat{V}(\zeta)(1 - \tau e^{-(1+\zeta)\tau}) &= (1 - \tau(1 + \zeta))e^{-(1+\zeta)\tau}, \\ \widehat{V}''(\zeta)(\zeta + e^{-(1+\zeta)\tau}) + 2\widehat{V}'(\zeta)(1 - \tau e^{-(1+\zeta)\tau}) \\ &\quad + \widehat{V}(\zeta)\tau^2 e^{-(1+\zeta)\tau} = -\tau(2 - \tau(1 + \zeta))e^{-(1+\zeta)\tau}, \end{aligned}$$

and by setting $\zeta = 0$, obtain

$$\begin{aligned} \widehat{V}(0)e^{-\tau} &= e^{-\tau}, \\ \widehat{V}'(0)e^{-\tau} + \widehat{V}(0)(1 - \tau e^{-\tau}) &= (1 - \tau)e^{-\tau}, \\ \widehat{V}''(0)e^{-\tau} + 2\widehat{V}'(0)(1 - \tau e^{-\tau}) + \widehat{V}(0)\tau^2 e^{-\tau} &= -\tau(2 - \tau)e^{-\tau}. \end{aligned}$$

By solving for the derivatives of \widehat{V} at the origin, we obtain $\widehat{V}(0) = 1$ (as expected), $\widehat{V}'(0) = -(e^\tau - 1)$, and $\widehat{V}''(0) = 2e^{2\tau}(1 - (1 + \tau)e^{-\tau})$. It follows that

$$\begin{aligned} \mathbb{E}(W_\tau) &= -\widehat{V}'(0) = e^\tau - 1, \\ \text{Var}(W_\tau) &= \widehat{V}''(0) - (-\widehat{V}'(0))^2 = e^{2\tau} - 2\tau e^\tau - 1. \end{aligned}$$

We may now quickly adapt the result to general exponential inter-renewal times by scaling. For any given $\lambda > 0$, the random variable $W_\tau^{(\lambda)} = \frac{1}{\lambda} W_{\lambda\tau}$ represents the waiting time for a gap of size τ for a Poisson renewal process of rate λ . The corresponding mean and variance of the scaled variable are hence given by

$$\mathbb{E}(W_\lambda^{(\tau)}) = \frac{e^{\lambda\tau} - 1}{\lambda}, \quad \text{Var}(W_\lambda^{(\tau)}) = \frac{e^{2\lambda\tau} - 2\lambda\tau e^{\lambda\tau} - 1}{\lambda^2}.$$

Of course, more detailed information is available through the distribution directly. Writing $V_\tau^{(\lambda)}(t)$ for the d.f. of $W_\tau^{(\lambda)}$ and, explicitly, $V(t) = V_\tau(t) = V_\tau^{(1)}(t)$ for the d.f. of $W_\tau = W_\tau^{(1)}$, we have $V_\tau^{(\lambda)}(t) = V_{\lambda\tau}(\lambda t)$ whence, by the scaling property of the transform, we also obtain $\widehat{V_\tau^{(\lambda)}}(\zeta) = \widehat{V_{\lambda\tau}}(\zeta/\lambda)$.

The model provides a framework for determining the duration before one may safely cross a traffic intersection—of admittedly dubious value for an impatient herd bent on crossing the road—or, for a traffic systems designer, a principled approach to determining the traffic signal duration patterns at an intersection.

5 Collective risk and the probability of ruin

An insurance company finds that claims arrive randomly at a mean rate of α per unit time, the number of claims up till time t governed by the Poisson process $N(t)$ of rate α . The claim amounts form a sequence of independent random variables X_1, X_2, \dots which are independent of the arrival process $N(t)$ and governed by a common probability law F with support in the positive half-line. We suppose that the distribution F has a finite mean μ which we may express as $\mu = \int_{\mathbb{R}^+} x dF(x) = \int_0^\infty [1 - F(x)] dx$ (see Problem XIII.3). The accumulated total payout made by the company up till time t is then given by $A(t) = X_1 + X_2 + \dots + X_{N(t)}$. By conditioning on the number of arrivals up till time t , the d.f. of the payout may be seen to be

$$\mathbb{P}\{A(t) \leq x\} = e^{-\alpha t} \sum_{n=0}^{\infty} \frac{(\alpha t)^n}{n!} F^{*n}(x).$$

A d.f. of this form is called a *compound Poisson distribution*.

To offset the payouts the company begins at time $t = 0$ with an initial reserve of capital ξ and continuously accumulates premiums at a fixed rate which we may, by a proper normalisation of the units of the time axis, consider to be unit.³ At any $t \geq 0$, the company's net cash reserve before payouts is hence

³There is no loss of generality in assuming that premiums accumulate continuously at a unit rate. If the premium rate is r per unit time and claims arrive at a mean rate α then a scaling of the time axis by the factor r results in a unit premium rate with the claim arrival rate becoming α/r in the normalised setting. All our expressions now go through with the replacement of α by α/r .

$\xi + t$. A sample path of the company's net cash reserve after payouts is shown in Figure 1. The company faces *ruin* if, at any time $t > 0$, its obligations exceed its reserve, that is to say, if the event $\{A(t) > \xi + t\}$ occurs. Considering the claim rate α as a fixed parameter, the probability of eventual ruin depends solely on the initial reserve ξ . Accordingly, let $R(\xi)$ denote the probability that, starting with an initial capital reserve ξ , the company *never* faces ruin. The function $R(\cdot)$ is a distribution function concentrated on the positive half-line and, aside from a possible jump of size $R(0)$ at the origin, we anticipate naturally enough that $R(\xi)$ increases continuously on $(0, \infty)$.

The nature of the chances of ruin (or, equivalently, solvency) is made clear by focusing on the first claim. Suppose the first claim occurs at $t = \tau$ in the amount $X_1 = y$. The company avoids ruin at that instant if, and only if, $y \leq \xi + \tau$. Assuming the company has not been ruined by the first claim, its reserve after the first payout has become $\xi + \tau - y$. The situation now resets with the new reserve and, as subsequent arrivals and payouts are independent of the first, the probability that the company never faces ruin going forward is $R(\xi + \tau - y)$. By integrating out with respect to the location and the amount of the first claim, we obtain

$$\begin{aligned} R(\xi) &= \int_0^\infty \left(\int_{[0, \xi+\tau]} R(\xi + \tau - y) dF(y) \right) \alpha e^{-\alpha\tau} d\tau \\ &= \alpha e^{\alpha\xi} \int_\xi^\infty \left(\int_{[0, \eta]} R(\eta - y) dF(y) \right) e^{-\alpha\eta} d\eta, \end{aligned} \quad (5.1)$$

the final form following by the change of variable $\xi + \tau = \eta$. It follows that R is differentiable and via the chain rule of differentiation we obtain

$$R'(\xi) = \alpha R(\xi) - \alpha \int_{[0, \xi]} R(\xi - y) dF(y).$$

We may wrangle the expression into a somewhat more familiar form by integrating over the interval $(0, x)$, resulting in

$$R(x) - R(0) = \alpha \int_0^x R(\xi) d\xi - \alpha \int_0^x \int_{[0, \xi]} R(\xi - y) dF(y) d\xi. \quad (5.2)$$

The integral in the second term on the right may be further simplified by reversing the order of integration to obtain

$$\int_0^x \int_{[0, \xi]} R(\xi - y) dF(y) d\xi = \int_{[0, x]} \int_y^x R(\xi - y) d\xi dF(y) = \int_0^x R(x - y) F(y) dy.$$

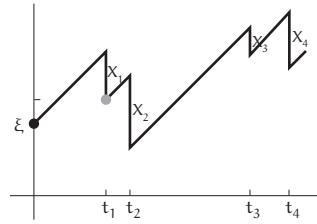


Figure 1: A sample path of the company's cash reserve.

[Set $u(y) = \int_y^x R(\xi - y) d\xi = \int_0^{x-y} R(z) dz$ and integrate by parts using Theorem XIV.5.5.] Manipulating the first term on the right of (5.2) by the change of variable $y = x - \xi$, we may write

$$\int_0^x R(\xi) d\xi = \int_0^x R(x - y) dy,$$

and, consequently, by a rearrangement of the terms in (5.2), we obtain

$$R(x) = R(0) + \alpha \int_0^x R(x - y)[1 - F(y)] dy \quad (x > 0).$$

This is the form we seek. *The probability of never facing ruin, $R(x)$, viewed as a function of the initial capital reserve x , is governed by the renewal equation*

$$R(x) = R(0)H_0(x) + (H * R)(x)$$

where $R(0)H_0(x)$ denotes the point measure of mass $R(0)$ concentrated at the origin and $H(x)$ denotes the (defective) distribution with density $H'(x) = \alpha(1 - F(x))$ concentrated on the positive half-line. As $R(\infty) = 1$, it is clear that $R(x)$ is a d.f. concentrated on the positive half-line. This d.f. has a jump of size $R(0)$ at the origin.

The convolutional relation suggests a passage to Laplace transforms of the renewal equation at hand and, as in (3.2), we obtain

$$\widehat{R}(\zeta) = \frac{R(0)\widehat{H}_0(\zeta)}{1 - \widehat{H}(\zeta)} \quad (\zeta > 0).$$

As $\widehat{H}_0(\zeta) = 1$ we only need to determine the Laplace transform of the defective distribution H . We have

$$\widehat{H}(\zeta) = \alpha \int_0^\infty e^{-\zeta x} (1 - F(x)) dx = \alpha \int_0^\infty e^{-\zeta x} dx - \alpha \int_0^\infty e^{-\zeta x} F(x) dx.$$

As an easy integration by parts of the Laplace transform of F shows that

$$\widehat{F}(\zeta) = \int_{[0, \infty)} e^{-\zeta x} dF(x) = e^{-\zeta x} F(x) \Big|_{0-}^\infty + \zeta \int_0^\infty e^{-\zeta x} F(x) dx,$$

it follows that

$$\widehat{H}(\zeta) = \alpha \left(\frac{1 - \widehat{F}(\zeta)}{\zeta} \right) \quad (\zeta > 0).$$

The Laplace transform of the distribution associated with the probability of never facing ruin is hence given by

$$\widehat{R}(\zeta) = \frac{R(0)}{1 - \alpha \left(\frac{1 - \widehat{F}(\zeta)}{\zeta} \right)} \quad (\zeta > 0),$$

and it only remains to determine the size of the jump of R at the origin. By allowing ζ to tend to zero, on the left we have $\widehat{R}(\zeta) \rightarrow 1$ as R is a proper distribution. On the other hand, as $\zeta \rightarrow 0$ on the right, the derivative property of the transform shows that

$$\frac{1 - \widehat{F}(\zeta)}{\zeta} = \frac{\widehat{F}(0) - \widehat{F}(\zeta)}{\zeta} \rightarrow -\widehat{F}'(0) = \int_{\mathbb{R}^+} x dF(x) = \mu. \quad (5.3)$$

The probability that ruin is avoided if no initial reserve is set aside is hence given by $R(0) = 1 - \alpha\mu$. This expression makes sense only if $\alpha\mu < 1$. The constraint is quite intuitive: as claims arrive at a mean rate α and the expected value of each claim is μ , the quantity $\alpha\mu$ represents the mean payout rate. If the company has set aside no initial reserve then, to avoid ruin, the payout stream must be covered by the premium stream which means that $\alpha\mu < 1$.

We have now obtained an elegant closed-form expression for the Laplace transform of R :

$$\widehat{R}(\zeta) = \frac{1 - \alpha\mu}{1 - \widehat{H}(\zeta)} = \frac{1 - \alpha\mu}{1 - \alpha \left(\frac{1 - \widehat{F}(\zeta)}{\zeta} \right)} \quad (\zeta > 0). \quad (5.4')$$

It is now not difficult to formally invert the Laplace transform and obtain an expression for $R(x)$. By expanding $(1 - \widehat{H}(\zeta))^{-1}$ in a geometric series, in view of the linearity and convolution properties of the transform we see that the expression may be inverted to obtain

$$R(x) = (1 - \alpha\mu) \sum_{n=0}^{\infty} H^{*n}(x) \quad (x > 0), \quad (5.4)$$

where H is the defective distribution with density $H'(x) = \alpha(1 - F(x))$ concentrated on the positive half-line. Even the weariest river winds somewhere safe to sea. Summarising our discoveries we obtain

THE POLLACZEK-KHINCHIN THEOREM *If $\alpha\mu < 1$, the probability of never facing ruin, $R(x)$, viewed as a function of the initial capital reserve x , and the Laplace transform $\widehat{R}(\zeta)$ of the associated distribution satisfy (5.4,5.4'). If $\alpha\mu > 1$ then $R(x) = 0$ for all x and ruin is certain whatever the initial capital reserve.*

The moments of R may be obtained by differentiation of \widehat{R} and we shall see explicit formulæ in the next section. But more can be said. The simple formulations (5.4, 5.4') may be used as a springboard for a more searching examination of the fine structure of ruin probabilities. While we shall not dig further into the theory in this volume, the following sample may whet the reader's appetite to read more on the theory of renewals. Suppose the tails of the payout distribution F are exponentially bounded and, more, that there exists κ such that

$$\alpha \int_{[0, \infty)} e^{\kappa x} [1 - F(x)] dx = 1 \text{ and } \mu^\# = \alpha \int_{[0, \infty)} x e^{\kappa x} [1 - F(x)] dx < \infty.$$

Then the probability of eventual ruin is given asymptotically by

$$1 - R(x) \sim \frac{1 - \alpha\mu}{\kappa\mu^\#} e^{-\kappa x}$$

as $x \rightarrow \infty$. The relation \sim is to be taken to mean that the ratio of the two sides tends to one as $x \rightarrow \infty$. This is Cramér's famous estimate for ruin.⁴

The pair of equations (5.4,5.4') are called the *Pollaczek–Khinchin formulæ*. While their derivation seems to be rather specialised to this particular context they are in reality in wide use, their ubiquitous appearance due to a basic connection with the theory of random walks.

Suppose the claims arrive sequentially at epochs t_1, t_2, \dots . Then the company faces ruin if $A(t_n) - t_n > \xi$ for any n . Suppose T_1, T_2, \dots represents the sequence of claim inter-arrival times, T_j representing the time between the $(j-1)$ th and j th claims. Here, the T_j are independent with a common exponential distribution of mean $1/\alpha$. Then $t_n = T_1 + \dots + T_n$ and, accordingly, the condition for ruin with the n th claim may be expressed in the form $\sum_{j=1}^n (X_j - T_j) > \xi$. The differences $U_j = X_j - T_j$ form a sequence of independent random variables with a common distribution whose mean $\mu - 1/\alpha$ we will naturally require to be negative, else the company quickly faces ruin. The partial sums $S_n = U_1 + \dots + U_n$ hence represent a *generalised random walk with a drift to $-\infty$* . Ruin probabilities are dictated by the maximum positive excursion $M = \max\{0, S_1, S_2, \dots\}$ of this random walk: more specifically, *starting with an initial capital reserve ξ , the company avoids ruin if, and only if, $S_n \leq \xi$ for every n , or, equivalently,*

$$R(\xi) = P\{M \leq \xi\} \quad (5.5)$$

is the d.f. of the maximal excursion M of the underlying random walk. This unexpected connection between collective risk and the maxima of random walks with a drift to $-\infty$ proves to be very profitable, the Pollaczek–Khinchin formulæ finding extensive use in queuing theory. We shall see why next.

6 The queuing process

A service station staffed by a single server is thrown open to the public at time $t = 0$ and customers begin to arrive in accordance with a Poisson process $N(t)$ of rate α . It will be convenient to label the arriving customers $n = 0, 1, 2, \dots$, and accordingly the inter-arrival times T_0, T_1, T_2, \dots of the customers form an

⁴For the original formulation see H. Cramér, "On some questions connected with mathematical risk", *University of California Publications in Statistics*, vol. 2, no. 5, pp. 99–125, 1954. Cramér's asymptotic estimates were originally obtained by deep complex variable methods. For a more elementary derivation starting from the renewal equation see W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, Wiley, New York, 1971.

independent sequence of random variables with a common exponential distribution of mean $1/\alpha$. The amount of service time needed by each customer is modelled by a sequence of independent positive random variables X_0, X_1, X_2, \dots , drawn from a common distribution F with support in the positive half-line, the sequence independent of the sequence of inter-arrival times. If, when the n th customer arrives she finds the server idle, she promptly gets attention from the server and departs after X_n units of time; if, on the other hand, she finds the server busy with other customers, she joins the queue and waits until the server has attended to all the preceding customers before she receives her X_n units of service.

For each n , let W_n denote the *waiting time* in queue of the n th customer; this is the duration from the moment of her arrival to the time she begins receiving service. The sequence $\{W_n, n \geq 0\}$ is called the *queuing process*. This process is best understood in a recursive formulation. Suppose that the n th customer arrives at time t_n , waits W_n units of time, receives X_n units of service, and departs at time $t_n + W_n + X_n$. The $(n+1)$ th customer, meanwhile, arrives at time $t_n + T_{n+1}$; she finds the server idle if $t_n + W_n + X_n < t_n + T_{n+1}$, in which case she receives immediate service without a wait; if $t_n + T_{n+1} \leq t_n + W_n + X_n$ on the other hand, she will find the server busy and will then wait for a duration $W_n + X_n - T_{n+1}$ before she receives service. Accordingly, we may compactly write the waiting time of the $(n+1)$ th customer in the recurrence

$$W_{n+1} = (W_n + X_n - T_{n+1})^+ \quad (n \geq 0),$$

where, as usual, the expression on the right denotes the positive part of $W_n + X_n - T_{n+1}$.

A feel for the salient features of the queuing process may be obtained by tracking its progress in time. The initial customer arrives at time T_0 and, finding no one has anticipated her in the queue, receives immediate service whence $W_0 = 0$. If the next customer arrives before service to the first customer is concluded, she faces a waiting time $W_1 = X_0 - T_1$. If the next arriving customer finds the server busy, she then waits a period $W_2 = W_1 + X_1 - T_2 = (X_0 - T_1) + (X_1 - T_2)$, and the process continues in this fashion until the *first* subsequent customer who arrives to find the server idle. At this point the process probabilistically restarts and it is clear that we have an imbedded renewal process with renewal epochs identified with a sequence of arrival indices $0 = \nu_0 < \nu_1 < \nu_1 + \nu_2 < \nu_1 + \nu_2 + \nu_3 < \dots$ at which arriving customers find the server idle. Each of these epochs punctuate *busy periods* for the server.

It will be convenient at this point to introduce the independent difference sequence $U_1 = X_0 - T_1, U_2 = X_1 - T_2, U_3 = X_2 - T_3, \dots$. The corresponding sequence of partial sums $S_n = U_1 + U_2 + \dots + U_n$ now represents a generalised random walk. By convention we also set $S_0 = 0$ to help unify expressions.

In this notation, for the first busy period, we begin with a renewal epoch $W_{\nu_0} = W_0 = 0$, and then proceed through a sequence of strictly positive

waiting times

$$\begin{aligned} W_1 &= W_0 + U_1 = S_1 - S_0, \\ W_2 &= W_1 + U_2 = S_2 - S_0, \\ \dots & \\ W_{v_1-1} &= W_{v_1-2} + U_{v_1-1} = S_{v_1-1} - S_0, \end{aligned}$$

until the process restarts with the v_1 th arrival. The next busy period is a probabilistic copy of the first with $W_{v_1} = 0$ followed by a sequence of strictly positive waiting times

$$\begin{aligned} W_{v_1+1} &= W_{v_1} + U_{v_1+1} = S_{v_1+1} - S_{v_1}, \\ W_{v_1+2} &= W_{v_1+1} + U_{v_1+2} = S_{v_1+2} - S_{v_1}, \\ \dots & \\ W_{v_1+v_2-1} &= W_{v_1+v_2-2} + U_{v_1+v_2-1} = S_{v_1+v_2-1} - S_{v_1} \end{aligned}$$

terminated by yet another renewal. The pattern is now clear. For the k th arrival during the j th busy period, we have $W_{v_{j-1}+k} = S_{v_{j-1}+k} - S_{v_{j-1}} > 0$ so that

$$S_{v_{j-1}+k} > S_{v_{j-1}} \quad (1 \leq k < v_j). \quad (6.1)$$

At the renewal epochs themselves, we have $W_{v_1+\dots+v_j} = 0$ so that $W_{v_{j-1}} + U_{v_j} = S_{v_1+\dots+v_j} - S_{v_1+\dots+v_{j-1}} \leq 0$. It follows that

$$0 \geq S_{v_1} \geq S_{v_2} \geq S_{v_3} \geq \dots \quad (6.1')$$

and the renewal epochs correspond to successively lower minima of the associated random walk S_n . These are called *ladder epochs* of the walk. A picture makes all clear, see Figure 2.

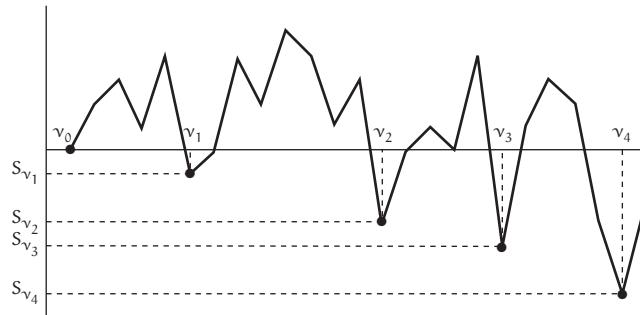


Figure 2: Ladder epochs of the queuing process.

For any arrival index n , let $\nu = \nu(n) \leq n$ be the last renewal index prior to (or equal to) n , that is, the ν th customer is the last prior (or current) customer to discover an idle server. Then

$$W_n = S_n - S_\nu = U_{\nu+1} + U_{\nu+2} + \dots + U_n.$$

As v is the last renewal index up till n it follows by (6.1,6.1') that $S_v \leq S_k$ for each $1 \leq k \leq n$ and, in consequence, $W_n = \max\{S_n - S_1, S_n - S_2, \dots, S_n - S_n\}$. A happy discovery of F. Pollaczek magically clarifies the setting.

Set $U'_1 = U_n$, $U'_2 = U_{n-1}, \dots, U'_{n-1} = U_2$, and $U'_n = U_1$, and consider the reversed walk $S'_k = U'_1 + \dots + U'_k$. It is then clear that $W_n = \max\{0, S'_1, \dots, S'_n\}$ and, as the reversed walk has exactly the same distribution as the original walk, we obtain a remarkable result.

THEOREM 1 *The waiting time W_n has the same distribution as $\max\{0, S_1, \dots, S_n\}$.*

By allowing $n \rightarrow \infty$ we see that the limiting waiting time distribution is exactly that of the maximum positive excursion $M = \max\{0, S_1, S_2, \dots\}$ of the underlying random walk. But this is a familiar setting: *the limiting distribution is exactly that of (5.5) corresponding to the probability that ruin is avoided in the setting of collective risk*. The results of the previous section may hence be imported wholesale into this new context.

THEOREM 2 *Suppose $\alpha\mu < 1$. Then, as $n \rightarrow \infty$, the distribution of the waiting time W_n tends to a limiting distribution R . This distribution and the associated Laplace transform \widehat{R} are given by the pair of Pollaczek–Khinchin formulae (5.4,5.4').*

The limiting distribution R may be identified with the waiting time distribution “in the steady state” after the queue has been operational for a long time.

Various moments may be deduced from (5.4') by differentiation. In particular, by clearing the denominator and taking derivatives of both sides, we obtain

$$\widehat{R}'(\zeta) \left(1 - \alpha \cdot \frac{1 - \widehat{F}(\zeta)}{\zeta} \right) - \alpha \widehat{R}(\zeta) \left(\frac{\widehat{F}(\zeta) - 1 - \zeta \widehat{F}'(\zeta)}{\zeta^2} \right) = 0 \quad (\zeta > 0). \quad (6.2)$$

By the derivative property (5.3), we have seen that $\zeta^{-1}[1 - \widehat{F}(\zeta)] \rightarrow -\widehat{F}'(0)$ as $\zeta \rightarrow 0$ provided F has expectation. Suppose now, additionally, that F has a finite second moment. By a Taylor expansion of \widehat{F} through two terms, we have $\widehat{F}(\zeta) = 1 + \zeta \widehat{F}'(\zeta) + \frac{1}{2}\zeta^2 \widehat{F}''(\eta)$ for some $\eta = \eta(\zeta)$ in the interval $(0, \zeta)$. As $\widehat{F}''(\zeta)$ is continuous, it follows that

$$\lim_{\zeta \downarrow 0} \frac{\widehat{F}(\zeta) - 1 - \zeta \widehat{F}'(\zeta)}{\zeta^2} = \frac{\widehat{F}''(0)}{2}.$$

By letting ζ decrease to zero in (6.2), we hence obtain

$$\widehat{R}'(0) + \frac{\alpha \widehat{F}''(0)}{2[1 + \alpha \widehat{F}'(0)]} = 0$$

and a compact formula for the mean waiting time stands revealed.

THEOREM 3 Suppose the random variable X has distribution F , the random variable W has distribution R , these variables representing a generic service time and a limiting waiting time, respectively. Suppose $E(X) < 1/\alpha$ and $E(X^2) < \infty$. Then W is integrable and has expectation

$$E(W) = \frac{\alpha E(X^2)}{2(1 - \alpha E(X))}.$$

This is called the *Pollaczek–Khinchin mean value formula*.

Writing $\mu = E(X)$ we may interpret $1/\mu$ as the *service rate* provided by a perpetually busy server. Consequently, $p_{\text{busy}} = \alpha\mu$ represents the fraction of time that the server is busy, with $p_{\text{idle}} = 1 - \alpha\mu$ the fraction of time that she is idle. Writing $\rho = p_{\text{busy}}/p_{\text{idle}}$, the rôle played by the distribution of service times becomes a little clearer if we write the mean value formula in the form $E(W) = \rho E(X^2)/2 E(X)$. If there is no variation in service times and $E(X) = \mu$ (that is to say, the service time distribution is equal to μ with probability one) then $E(X^2) = \mu^2$ and it follows that $E(W) = \rho\mu/2$. As $E(X^2) = \text{Var}(X) + E(X)^2 \geq \mu^2$ with equality if, and only if, $\text{Var}(X) = 0$, it follows that *among all service distributions of a given mean μ , deterministic service sees the smallest expected waiting time*. In particular, if X is exponentially distributed with mean μ then $E(X^2) = 2\mu^2$ and $E(W) = \rho\mu$ so that the expected waiting time for exponential service is twice as long as that for deterministic service. This is a consequence of the *inspection paradox* discussed at length in Section IX.7: from the vantage point of an arriving customer, the residual service time of a customer in service is atypical with longer service times more likely to be interrupted.

7 Ladder indices and a combinatorial digression

We have seen general random walks appear imbedded in applications in collective risk and queuing in the previous two sections. These random walks were characterised by a drift to $-\infty$. Rather different behaviours emerge when the walk has no systemic bias: the basic example is that of the simple random walk discussed in Sections VIII.4,5. The reader who has explored these sections will recall the quite amazing properties of fluctuations that emerged but may, understandably, given the context, have been led to feel that these remarkable properties are somehow intimately tied to the particular nature of the binomial distribution. The discovery that these laws of fluctuations actually held much more generally occasioned great surprise when they were first discovered but the deep and arduous method of proof provided little intuition as to why this was so. It was only by a happy discovery of E. Sparre Andersen that simple combinatorial structures were at the root of these general laws that clarity was brought to the matter.

Sparre Andersen's combinatorial methods were gradually simplified by many authors until the simple combinatorial lemma⁵ that was at the heart of the matter was exposed. It will be simplest if we begin with it here before returning to random walks in the next section.

Given n and a finite sequence of real numbers x_1, \dots, x_n , we form the partial sums $s_0 = 0, s_1 = x_1, s_2 = x_1 + x_2, \dots, s_n = x_1 + \dots + x_n$.

DEFINITION An integer $\nu \in \{1, \dots, n\}$ is a *ladder index* for the sequence x_1, \dots, x_n if $s_\nu > s_0, s_\nu > s_1, \dots, s_\nu > s_{\nu-1}$. We say that the sequence has a *terminal ladder index* if n is a ladder index.

In a more precise terminology we should associate ladder indices with the partial sums s_0, s_1, \dots, s_n of the given sequence but there is no danger of ambiguity as the partial sums uniquely determine the underlying sequence x_1, \dots, x_n .

A sequence of length n may have 0, 1, 2, ..., or n ladder indices. Extreme examples are provided by sequences all of whose elements are strictly negative or all strictly positive. In the former case the partial sums are strictly decreasing, $0 = s_0 > s_1 > \dots > s_n$, and there are no ladder indices; in the latter case the partial sums are strictly increasing, $0 = s_0 < s_1 < \dots < s_n$, and each of the numbers 1, 2, ..., n is a ladder index. A sufficient (but not necessary) condition for the sequence to have at least one ladder index is that $s_n > 0$. (The reader should convince herself why this is true.)

EXAMPLE 1 The length 4 sequence $-1, 1, -1, 1$ has no ladder indices while the reversed sequence $1, -1, 1, -1$ has exactly one ladder index 1. The length 7 sequence $-2, 5, 3, -6, 10, -1, 1$ has sum $s_7 = 10$ and hence has at least one ladder index; in fact it has three ladder indices 2, 3, and 5. The length 7 sequence $-2, 1, -5, 4, 2, -7, 10$ has exactly one ladder index which is terminal. ►

Ladder indices identify locations where the running total of partial sums attains a maximum up to that point. Of course, a completely analogous theory can be built up for minima instead. We may distinguish between the cases by adding the qualifiers "ascending" and "descending" to the ladder indices but it will hardly be necessary for our purposes and, for definiteness, we will stick with ladder indices of the ascending type.

Starting with the given sequence $C^{(0)} = x_1, x_2, \dots, x_n$ we may create $n - 1$ additional cyclical arrangements

$$C^{(\nu)} = x_{\nu+1}, x_{\nu+2}, \dots, x_n, x_1, x_2, \dots, x_\nu \quad (1 \leq \nu \leq n - 1),$$

where it is convenient to number the cyclical arrangements from 0 through $n - 1$ beginning with the given sequence and each successive sequence identified by

⁵W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, op. cit. Chapter XII fleshes out the theory of random walks on the line in detail. The combinatorial lemma referred to here can be found in Section XII.6.

the last element. (If we identify 0 with n then this interpretation holds also for the original sequence.) For each $\nu \in \{0, 1, \dots, n - 1\}$, the ν th cyclical arrangement $C^{(\nu)}$ engenders a corresponding sequence of partial sums $0 = s_0^{(\nu)}, s_1^{(\nu)}, \dots, s_n^{(\nu)}$ which we may relate to the original partial sums via

$$s_k^{(\nu)} = \begin{cases} s_{\nu+k} - s_\nu & \text{if } 1 \leq k \leq n - \nu, \\ s_n - s_\nu + s_{k-n+\nu} & \text{if } n - \nu + 1 \leq k \leq n. \end{cases} \quad (7.1)$$

The second line of the equation shows that, as expected, $s_n^{(\nu)} = s_n$ for each ν , the final member of the partial sums of an arrangement being simply the sum of all the terms of the sequence.

A consideration of cyclical arrangements helps symmetrise the vagaries of a given sequence but the symmetry is not apparent. The key is to consider those cyclical arrangements which have a terminal ladder index.

A COMBINATORIAL LEMMA *There exists a cyclical arrangement with a terminal ladder index if, and only if, $s_n > 0$. Moreover, if, for some $r \geq 1$, there is a cyclical arrangement with a terminal ladder index for which n is the r th ladder index, then there exist precisely r cyclical arrangements with a terminal ladder index and each of these arrangements contains exactly r ladder indices.*

PROOF: If $s_n < 0$ then $s_n^{(\nu)} = s_n < 0$ for each ν and no circular arrangement can have a terminal ladder index. Suppose hence that $s_n > 0$. As $s_0 = 0$, there exists a maximal partial sum. To take care of the possibility that the maximum may be repeatedly achieved, we focus on the *first* index ν at which it is achieved,

$$s_\nu > s_0, s_\nu > s_1, \dots, s_\nu > s_{\nu-1} \text{ and } s_\nu \geq s_{\nu+1}, s_\nu \geq s_{\nu+2}, \dots, s_\nu \geq s_n.$$

Then for the cyclical arrangement $C^{(\nu)}$ by (7.1) we have

$$\begin{aligned} s_k^{(\nu)} &\leq 0 && \text{if } 1 \leq k \leq n - \nu, \\ s_k^{(\nu)} &< s_n && \text{if } n - \nu + 1 \leq k \leq n - 1, \end{aligned}$$

and thus $C^{(\nu)}$ is an arrangement with a terminal ladder index whence there exists at least one such arrangement. By renumbering the cyclical arrangements starting with this one we may now, without loss of generality, suppose that the original sequence $C^{(0)}$ has a terminal ladder index.

Suppose accordingly that the given sequence $C^{(0)}$ has r ladder indices, n being the last. As $s_n^{(\nu)} = s_n$ for each ν , the second line of (7.1) shows then that for $n - \nu + 1 \leq k \leq n - 1$, the inequality $s_k^{(\nu)} < s_n^{(\nu)} = s_n$ is strict if, and only if, $s_\nu > s_1, \dots, s_\nu > s_{\nu-1}$, that is to say, if ν is a ladder index of $C^{(0)}$. Thus, the only candidates for terminal ladder indices are those sequences $C^{(\nu)}$ for which ν is a ladder index of the sequence $C^{(0)}$. If ν is a ladder index of

$C^{(0)}$, the corresponding partial sum must be strictly positive, $s_\nu > 0$; moreover, s_n is the maximal element of the partial sums and so $s_n > s_0, s_n > s_1, \dots, s_n > s_{n-1}$. It follows by the first line of (7.1) that for $1 \leq k \leq n - \nu$, we have $s_k^{(\nu)} = s_{\nu+k} - s_\nu < s_n - 0 = s_n^{(\nu)}$ whenever ν is a ladder index of the original sequence. Thus, n is a terminal ladder index precisely for those cyclical arrangements corresponding to ladder indices and so there are exactly r such sequences.

Finally, suppose that $1 \leq \nu_1, \dots, \nu_r = n$ are the ladder indices of the original sequence. Then $0 < s_{\nu_1} < s_{\nu_2} < \dots < s_{\nu_r} = s_n$. But then $s_{\nu_{j+1}} - s_{\nu_j} > 0$ for each $j \geq 1$ and, in consequence, each cyclical shift by a ladder index maintains the relative spacing between successive ladder indices as may be seen from the first line of (7.1). (The reader should imagine the sequence wrapping around with n adjacent to 1.) It follows that each of the cyclical arrangements $C^{(\nu_j)}$ has exactly r ladder indices. ▶

The reader may well find Figure 3 as persuasive as the formal argument given above. Fleshing out the simple example graphed in the figure may serve

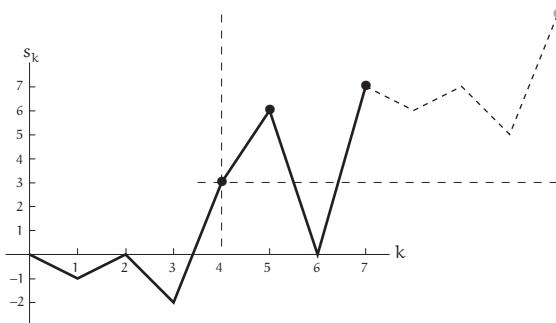


Figure 3: The graph of the partial sums of the sequence $C^{(0)} = -1, 1, -2, 5, 3, -6, 7$ has three ladder indices 4, 5, and 7 indicated by bullets. The graph of the partial sums of the cyclical arrangement $C^{(4)} = 3, -6, 7, -1, 1, -2, 5$ corresponding to a cyclical shift to the first ladder index is obtained by shifting the origin to the point (4, 3) corresponding to the coordinates of the first ladder index and attaching a copy of the first segment of the graph to its tail (dotted line) to obtain the new graph with corresponding ladder indices 1, 3, and 7.

to add even more verisimilitude to the bald narrative.

EXAMPLE 2) The sequence $-1, 1, -2, 5, 3, -6, 7$ has a total of three ladder indices, one of them terminal. The cyclical arrangements of the sequence and the corresponding partial sums (leaving out the common value $s_0 = 0$) are shown in Table 2. As asseverated, there are three cyclical arrangements with terminal ladder indices ($C^{(0)}, C^{(4)}$, and $C^{(5)}$), and each of these has three ladder indices. While other cyclical arrangements ($C^{(2)}, C^{(3)}$, and $C^{(6)}$) also have three ladder indices, none of them has a terminal ladder index. The reader

	Cyclical arrangements							Partial sums							
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	s_1	s_2	s_3	s_4	s_5	s_6	s_7	r
$C^{(0)}$	-1	1	-2	5	3	-6	7	-1	0	-2	3	6	0	7	3
$C^{(1)}$	1	-2	5	3	-6	7	-1	• 1	-1	• 4	• 7	1	• 8	7	4
$C^{(2)}$	-2	5	3	-6	7	-1	1	-2	• 3	• 6	0	• 7	6	7	3
$C^{(3)}$	5	3	-6	7	-1	1	-2	• 5	• 8	2	• 9	8	9	7	3
$C^{(4)}$	3	-6	7	-1	1	-2	5	• 3	-3	• 4	3	4	2	• 7	3
$C^{(5)}$	-6	7	-1	1	-2	5	3	-6	• 1	0	1	-1	• 4	• 7	3
$C^{(6)}$	7	-1	1	-2	5	3	-6	• 7	6	7	5	• 10	• 13	7	3

Table 2: Cyclical arrangements and their partial sums. The locations of the ladder indices are marked with bullets.

should also note that $C^{(1)}$ has four ladder indices—cyclical arrangements do not, in general, have the same number of ladder indices. Those with terminal ladder indices form a special, regular subclass. ▶

Armed with this combinatorial observation we now turn to a consideration of the maxima of unbiased random walks.

8 The amazing properties of fluctuations

Suppose X_1, \dots, X_n, \dots is a sequence of random variables drawn by independent sampling according to a d.f. F with support on both the positive and negative portions of the line. We call the corresponding sequence of partial sums $S_0 = 0, S_1 = X_1, S_2 = X_1 + X_2, \dots, S_n = X_1 + \dots + X_n, \dots$ a *random walk on the line*. We turn now to the case when the walk is unbiased.

We recall that a d.f. F is *symmetric* if $F(x) = 1 - F(-x)$ at all points of continuity of F . The technical nuisance of a random walk repeatedly hitting a certain value is obviated if F is continuous and, accordingly, to keep the salient points in view, we suppose in this section that F is symmetric and continuous. If F has expectation then it must perforce be zero, $\int_{\mathbb{R}} x dF(x) = 0$, the standard normal distribution providing a typical example. A symmetric and continuous distribution may not have expectation, however, as evidenced by the Cauchy distribution.

By reasons of symmetry, the partial sum S_n has d.f. F^{*n} which is continuous and symmetric (though it may not have expectation). In consequence, $P\{S_n \leq 0\} = P\{S_n > 0\} = 1/2$ for each n and the walk is equally likely to be on either side of the abscissa.

What is the largest excursion one can anticipate over any finite segment S_0, S_1, \dots, S_n of the walk? This is surely a very natural question of interest. The

maximum value attained by the walk over any finite segment is built up on a scaffolding of peaks and valleys, each succeeding peak increasing the maximum value thus far attained. Each previous summit scaled resets the conditions of the walk, the next summit lying above the previously attained height. These peaks correspond to the ladder indices of the walk and we begin accordingly with a consideration of the distribution of ladder indices.

If there is to be a maximum at all, there must surely be a *first* summit where the walk first encroaches into positive territory. For each $n > 0$, let $f(n)$ denote the probability that S_n becomes strictly positive for the first time,

$$f(n) = \mathbf{P}\{S_1 \leq 0, S_2 \leq 0, \dots, S_{n-1} \leq 0, S_n > 0\}.$$

The values $\{f(n), n \geq 1\}$ determine the distribution of the first ladder index. (We shall see shortly that the distribution is *proper*, that is, $\sum_n f(n) = 1$.) Suppose now that v_1, v_2, \dots represent successive ladder indices of the walk. Setting $v_0 = 0$, it follows that $v_1 = v_1 - v_0$ has the distribution $\{f(n)\}$ and is the first instant at which $S_{v_1} > 0$. From the vantage point of the first ladder index [that is to say, with origin reset to the point (v_1, S_{v_1})] the next ladder index corresponds to the first succeeding epoch at which $S_{v_2} - S_{v_1} > 0$. As the values $S_{v_1+n} - S_{v_1} = X_{v_1+1} + \dots + X_{v_1+n}$ represent a new random walk independent of the value $S_{v_1} = X_1 + \dots + X_{v_1}$, it follows that $v_2 - v_1$ is independent of v_1 and has the same distribution $\{f(n)\}$. Proceeding in this fashion, the r th ladder index $v_r = (v_r - v_{r-1}) + (v_{r-1} - v_{r-2}) + \dots + (v_2 - v_1) + (v_1 - v_0)$ is the sum of r independent random variables with the common distribution $\{f(n), n > 1\}$. It follows that, for each $r \geq 1$, the ladder index v_r has the distribution $f^{*r}(n)$ given by the n -fold convolution of the arithmetic distribution f with itself.

The convolutional form of the distribution of ladder indices suggests a passage to Laplace transforms. Let $\mathfrak{F}(s) = \sum_{n=1}^{\infty} f(n)s^n$ be the generating function of the first ladder index, that is to say, of the distribution $\{f(n)\}$. By the convolution property of the Laplace transform, the generating function of the r th ladder index, that is to say, of the distribution $\{f^{*r}(n)\}$, is then given by $\mathfrak{F}(s)^r$.

Consider now the setting where, for any given n , the finite sequence X_1, \dots, X_n has a terminal ladder index. Then n is the r th ladder index of the sequence for some r . With n and r fixed, let Z be the indicator for this event, that is, Z is a Bernoulli variable taking value 1 if n is the r th ladder index of the sequence X_1, \dots, X_n , and value 0 otherwise. Then Z has success probability $f^{*r}(n)$ and it follows that $E(Z) = f^{*r}(n)$.

The combinatorial lemma of the previous section suggests that it would be profitable to consider those cyclical arrangements of the sequence X_1, \dots, X_n with terminal ladder indices. In the notation of that section, we let

$$C^{(0)} = X_1, X_2, \dots, X_n,$$

$$C^{(v)} = X_{v+1}, X_{v+2}, \dots, X_n, X_1, X_2, \dots, X_v \quad (1 \leq v \leq n-1)$$

represent successive cyclical arrangements of the original sequence. By the symmetry inherent in the situation, each cyclical arrangement $C^{(v)}$ has exactly the same distribution.

Let $Z^{(v)}$ now be the indicator that $C^{(v)}$ has a terminal r th ladder index; of course, in this expanded notation, $Z^{(0)} = Z$. Then each $Z^{(v)}$ is a Bernoulli variable with the common success probability $f^{*r}(n)$. Consequently,

$$\mathbb{E}(Z) = \frac{1}{n} [\mathbb{E}(Z^{(0)}) + \mathbb{E}(Z^{(1)}) + \cdots + \mathbb{E}(Z^{(n-1)})] = \frac{1}{n} \mathbb{E}(Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)})$$

by additivity of expectation. (The reader should note how elegantly additivity finesse the fact that while the variables $Z^{(0)}, Z^{(1)}, \dots, Z^{(n-1)}$ have a common marginal distribution, they are manifestly *not* independent.) The sum $Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)}$ represents the number of cyclical arrangements with a terminal r th ladder index. By the combinatorial lemma of the previous section, if there is one such cyclical arrangement there must be exactly r and, consequently, $Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)}$ can take values 0 or r only. It follows that

$$f^{*r}(n) = \frac{r}{n} \mathbf{P}\{Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)} = r\}.$$

Dividing both sides by r and summing over all possible values of r , we obtain

$$\sum_{r=1}^{\infty} \frac{f^{*r}(n)}{r} = \frac{1}{n} \sum_{r=1}^{\infty} \mathbf{P}\{Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)} = r\}.$$

(While it is true that only the terms $1 \leq r \leq n$ contribute to the sum, it does no harm to make the upper limit infinite as $f^{*r}(n) = 0$ if $r > n$. This saves a pernickety fiddling around with limits when we change the order of summation.) Again by our combinatorial lemma, there exists a cyclical arrangement of X_1, \dots, X_n with a terminal ladder index if, and only if, $S_n > 0$. And if n is a ladder index of some cyclical arrangement then it must be the r th ladder index for some $r \geq 1$, these events being mutually exclusive. It follows by countable additivity of probability measure that

$$\sum_{r=1}^{\infty} \mathbf{P}\{Z^{(0)} + Z^{(1)} + \cdots + Z^{(n-1)} = r\} = \mathbf{P}\{S_n > 0\} = \frac{1}{2}. \quad (8.1)$$

We hence obtain the very simple expression

$$\sum_{r=1}^{\infty} \frac{f^{*r}(n)}{r} = \frac{1}{2n}.$$

Multiplying both sides by s^n and summing over $n \geq 1$ now results in

$$\sum_{n=1}^{\infty} \sum_{r=1}^{\infty} \frac{f^{*r}(n)s^n}{r} = \frac{1}{2} \sum_{n=1}^{\infty} \frac{s^n}{n}.$$

If $|s| < 1$ there is no difficulty with the interchange of the order of summation on the left (we are dealing with a convergent series of positive terms) and so

$$\sum_{n=1}^{\infty} \sum_{r=1}^{\infty} \frac{f^{*r}(n)s^n}{r} = \sum_{r=1}^{\infty} \frac{1}{r} \sum_{n=1}^{\infty} f^{*r}(n)s^n = \sum_{r=1}^{\infty} \frac{\mathfrak{F}(s)^r}{r}$$

as, for each r , we recognise in the inner sum the generating function of the distribution $\{f^{*r}(n), n \geq 1\}$ [the reader should bear in mind that $f^{*r}(0) = 0$]. It follows that

$$\sum_{r=1}^{\infty} \frac{\mathfrak{F}(s)^r}{r} = \frac{1}{2} \sum_{n=1}^{\infty} \frac{s^n}{n},$$

these expressions convergent for all $|s| < 1$. Both expressions may be related to logarithms via the Taylor series expansion for the (natural) logarithm,

$$\log(1-x) = -x - \frac{1}{2}x^2 - \frac{1}{3}x^3 - \dots \quad (|x| < 1),$$

to obtain the remarkably simple expression $-\log(1 - \mathfrak{F}(s)) = -\frac{1}{2} \log(1 - s)$. Taking exponentials of both sides produces a beautifully compact expression.

THEOREM 1 *The generating function of the distribution of the first ladder index satisfies $\mathfrak{F}(s) = 1 - \sqrt{1-s}$ if the underlying d.f. F is symmetric and continuous.*

By expanding $1 - \sqrt{1-s}$ in a Taylor series (see Problems I.1–5) we may read out the probabilities $f(n)$ explicitly but we have other fish to fry. In passing, the reader should observe that $\sum_n f(n) = \mathfrak{F}(1) = 1$ and the first ladder index distribution is proper as asserted.

Let $u(n) = P\{S_n > S_0, S_n > S_1, \dots, S_n > S_{n-1}\}$ denote the probability that n is a ladder index (not necessarily the first). Let $\mathfrak{U}(s) = \sum_{n=0}^{\infty} u(n)s^n$ denote the corresponding generating function.

By conditioning on the epoch of occurrence of the first ladder index we see that, for $n \geq 1$, $u(n) = \sum_k u(n-k)f(k) = (f * u)(n)$, the summation only formally infinite in view of our conventions $f(j) = 0$ for $j \leq 0$ and $u(j) = 0$ for $j < 0$. The occurrence of ladder indices hence forms an arithmetic renewal process and we see that the renewal probabilities are governed by the discrete renewal equation (3.4). By proceeding to generating functions, (3.4') provides the fundamental relation $\mathfrak{U}(s) = [1 - \mathfrak{F}(s)]^{-1}$ linking the generating function of ladder indices to the generating function of the first ladder index. In view of Theorem 1, we obtain

$$\mathfrak{U}(s) = (1-s)^{-1/2} = \sum_{n=0}^{\infty} \binom{-1/2}{n} (-s)^n = \sum_{n=0}^{\infty} \binom{2n}{n} 2^{-2n} s^n$$

by expanding out $(1-s)^{-1/2}$ in a power series (see Problems I.1–5). By comparing terms with the formal power series $\mathfrak{U}(s) = \sum_{n=0}^{\infty} u(n)s^n$, we obtain a remarkably explicit formula for the ladder index distribution.

THEOREM 2 *The generating function of the distribution $\{u(n)\}$ of ladder indices satisfies $U(s) = (1-s)^{-1/2}$ if F is symmetric and continuous. The associated ladder index probabilities are given by*

$$u(n) = (-1)^n \binom{-1/2}{n} = \binom{2n}{n} 2^{-2n} \quad (n \geq 0).$$

The ladder index probabilities $u(n)$ are identical to the probabilities u_{2n} given in (VIII.5.1) that a simple random walk returns to the origin in $2n$ steps!

Pollaczek's device of the reversed walk that was introduced in Section VIII.5 proves just as efficacious in the general context. Set $X'_1 = X_n, X'_2 = X_{n-1}, \dots, X'_n = X_1$ and for $1 \leq k \leq n$ let $S'_k = X'_1 + \dots + X'_k$ denote the corresponding random walk. Then the inequalities $S_n > S_0, S_n > S_1, \dots, S_n > S_{n-1}$ are equivalent to $S'_n > 0, S'_{n-1} > 0, \dots, S'_1 > 0$. As the joint distribution of the variables (X_1, \dots, X_n) is invariant under permutation, it follows that the walks $\{S_k\}$ and $\{S'_k\}$ have the same distribution. We consequently rederive the fundamental identity

$$\mathbf{P}\{S_1 > 0, S_2 > 0, \dots, S_n > 0\} = \mathbf{P}\{S_n > S_0, S_n > S_1, \dots, S_n > S_{n-1}\} = u(n).$$

By reasons of symmetry we also obtain the companion identity

$$\mathbf{P}\{S_1 \leq 0, S_2 \leq 0, \dots, S_n \leq 0\} = \mathbf{P}\{S_n \leq S_0, S_n \leq S_1, \dots, S_n \leq S_{n-1}\} = u(n)$$

as the underlying distribution F is continuous and symmetric. (If the reader finds the appeal to symmetry specious she may wish to repeat the argument leading to (8.1) with the inequalities $>$ replaced by \leq and the replacement of the ascending ladder indices by descending ladder indices.)

With these identities in hand, we may now repeat the analysis of the maximum of the simple random walk almost verbatim. Indeed, let M_n denote the index m where the walk over n steps *first* achieves a maximum, that is to say, the inequalities

$$\begin{aligned} S_m &> S_0, S_m > S_1, \dots, S_m > S_{m-1}, \\ S_m &\geq S_{m+1}, S_m \geq S_{m+2}, \dots, S_m \geq S_n \end{aligned}$$

all hold. Then the analysis leading from (VIII.5.3) to (VIII.5.4) holds *in toto* [indeed, as the reader should verify, all that is needed is the independence of the X_j] and we obtain

$$\mathbf{P}\{M_n = m\} = \mathbf{P}\{S_1 > 0, \dots, S_m > 0\} \cdot \mathbf{P}\{S_1 \leq 0, \dots, S_{n-m} \leq 0\} = u(m)u(n-m).$$

In view of the formal equivalence of $u(n)$ and u_{2n} via Theorem 2, the situation is now completely analogous to that of (VIII.5.6) and leads to the same crashing conclusion.

THEOREM 3 Suppose F is symmetric and continuous, and suppose $0 < t < 1$. Then the index M_n of the maximum of the associated random walk over n steps satisfies the arc sine limit law $P\{M_n \leq tn\} \rightarrow \frac{2}{\pi} \arcsin \sqrt{t}$ as $n \rightarrow \infty$.

The surprising nature of this result was dealt with in detail in Section VIII.5.

The reader who is curious about strengthening the result will, on examination of the proof, find that the only place the continuity of F is required is in the assertion $P\{S_n > 0\} = P\{S_n \leq 0\} = 1/2$ used in (8.1). She should find it plausible that the limiting asymptotic results will continue to hold as long as $|P\{S_n > 0\} - 1/2|$ decreases to zero sufficiently fast. Sparre Andersen showed indeed that this is true. His *theorem*: *the arc sine limit law for maxima holds if the series $\sum_{n=1}^{\infty} \frac{1}{n} [P\{S_n > 0\} - \frac{1}{2}] = c$ converges.* F. Spitzer showed that it suffices for F to have zero expectation and finite variance for the series to converge and so the arc sine law applies widely to general symmetric distributions. As one might anticipate, squeezing the last drop out of the result requires a correspondingly deep technical armoury—in this case, deep Tauberian theorems—and I will not attempt to build up the necessary background here. Problem 18 provides more evidence in support of the Sparre Andersen theorem.

9 Pólya walks the walk

A simple random walk in two dimensions starts at the origin of the coordinate plane and at each epoch takes a unit step right, left, up, or down, each with equal probability $1/4$. The position of the walk at epoch n is now a lattice point S_n whose coordinates are given by a pair of integer values (k, l) . In three dimensions, likewise, the position of a walk is a lattice point S_n whose coordinates are given by a triple of integer values (k, l, m) : starting from its current position, at each epoch the walk takes a unit step in either the positive or the negative direction along any of the three coordinate axes, each of the six possible steps having equal probability $1/6$. The situation readily generalises to random walks in d dimensions. The walk begins at the coordinate origin and, at each step, takes a unit positive or negative step along any of the d coordinate axes, each of the $2d$ possible steps having equal probability $1/2d$. The position of the walk at any epoch n is then a lattice point S_n whose coordinates are given by a d -tuple of integer values (k_1, \dots, k_d) . What is the probability that the walk will return to the origin?

We reuse notation from Section VIII.5 and write u_{2n} for the probability of a return to the coordinate origin at step $2n$ with the natural boundary conditions $u_0 = 1$ and $u_{2n} = 0$ for $n < 0$. Likewise, we write f_{2n} for the probability of a *first* return to the origin at step $2n$; the natural boundary conditions are $f_{2n} = 0$ for $n \leq 0$. As a return to the origin in $2n$ steps necessarily implies a first return to the origin, by conditioning on a first return at step $2k$, we have $u_{2n} = \sum_k f_{2k} u_{2n-2k}$ for $n \geq 0$. The convolutional relationship suggests that a move to Laplace transforms may be profitable.

Let $\mathfrak{U}(s) = \sum_n u_{2n} s^{2n}$ and $\mathfrak{F}(s) = \sum_n f_{2n} s^{2n}$ be the generating functions of $\{u_{2n}\}$ and $\{f_{2n}\}$, respectively. Multiplying both sides of the recurrence by s^{2n} and summing over the entire range of validity $n \neq 0$ we obtain $\mathfrak{U}(s) - 1 = \mathfrak{F}(s)\mathfrak{U}(s)$, or, $\mathfrak{U}(s)\{1 - \mathfrak{F}(s)\} = 1$. It is clear by comparison with a geometric series that $\mathfrak{U}(s)$ and $\mathfrak{F}(s)$

are both convergent for $|s| < 1$. As $\sum_n f_{2n} = \mathfrak{F}(1)$ connotes the probability of an eventual return to the origin, we have $\mathfrak{F}(1) \leq 1$, and *a fortiori* the series $\sum_n f_{2n}$ converges. On the other hand, for every positive integer N , we have $\sum_{n=0}^N u_{2n} \leq \lim_{s \rightarrow 1} \mathfrak{U}(s) \leq \sum_{n=0}^{\infty} u_{2n}$ whence the frequency of returns to the origin is governed by the behaviour of the series $\sum_n u_{2n}$.

THEOREM 1 *A random walk on a lattice returns to the origin with probability one if, and only if, the series $\sum_n u_{2n}$ is divergent.*

PROOF: If the probability of return to the origin is less than one then $\mathfrak{F}(1) = f < 1$. It follows that the series $\mathfrak{U}(1) = \sum_n u_{2n}$ is convergent and equal to $1/(1-f)$. On the other hand, if the walk returns to the origin with probability one then $\mathfrak{F}(1) = 1$. It follows then that $\lim_{s \rightarrow 1} \mathfrak{U}(s) = \infty$ whence the series $\sum_n u_{2n}$ diverges. ▶

Let us immediately apply this observation to walks in one, two, and three dimensions. In one dimension, $u_{2n} = \binom{2n}{n} 2^{-2n}$; the expression on the right is the central term of the binomial which is asymptotic to $(\pi n)^{-1/2}$ [see (XIV.7.4)]. The series $\sum_n n^{-1/2}$ diverges and the walk, in consequence, returns to the origin with probability one, as we've already seen in Theorem VIII.5.3.

A walk in two dimensions will return to the origin if, and only if, the number of positive steps in each of the dimensions is exactly counterbalanced by an equal number of negative steps in that dimension. Over $2n$ steps, the number of paths with j right, j left, $n-j$ up, and $n-j$ down steps is $\binom{2n}{j} \binom{2n-j}{j} \binom{2n-2j}{n-j} \binom{n-j}{n-j}$ which, after simplification and collection of terms, becomes $\binom{2n}{n} \binom{n}{j}^2$. It follows that $u_{2n} = 4^{-2n} \binom{2n}{n} \sum_j \binom{n}{j}^2$. The sum on the right is identically $\binom{2n}{n}$ (see Problem VIII.2) so that $u_{2n} = [2^{-2n} \binom{2n}{n}]^2 \sim \frac{1}{\pi n}$ by another application of the asymptotic estimate (XIV.7.4) of the central term of the binomial. The harmonic series $\sum_n 1/n$ diverges as we know from elementary calculus (or see Problem XIV.30) and the series $\sum_n u_{2n}$ is again divergent.

Arguing in this fashion, in three dimensions a return to the origin obtains when the positive steps are cancelled by an equal number of negative steps in each dimension whence we obtain

$$u_{2n} = \sum_{j,k} \frac{(2n)!}{j!k!(n-j-k)!^2} 6^{-2n} = \binom{2n}{n} 2^{-2n} \sum_{j,k} \left(\frac{n!}{j!k!(n-j-k)!} 3^{-n} \right)^2 \quad (9.1)$$

where the sum is over positive indices j and k with $j+k \leq n$. Within the round brackets on the right we identify the terms of a trinomial distribution and writing $t_n = \max_{j,k} 3^{-n} n! / (j!k!(n-j-k)!)$ for its maximum value, we may bound

$$\sum_{j,k} \left(\frac{n!}{j!k!(n-j-k)!} 3^{-n} \right)^2 \leq t_n \sum_{j,k} \frac{n!}{j!k!(n-j-k)!} 3^{-n} = t_n$$

as the sum over all terms of the trinomial distribution must add to one (Problem VIII.5). The sum of squares on the right in (9.1) is hence bounded above by the largest term of the trinomial distribution. Having identified where we should look it is now not difficult to finish off the proof. By an elementary combinatorial argument it is not difficult to see that the trinomial $3^{-n} n! / (j!k!(n-j-k)!)$ achieves its maximum when j , k , and $n-j-k$ are

all in the immediate vicinity of $n/3$ (the reader who would like a pointer should consult Problems VIII.6,7). A repeated application of Stirling's formula (XIV.7.3) now shows that the maximum term of the trinomial is asymptotic to $t_n \sim 3^{3/2}/(2\pi n)$. Coupling this with our estimate for the central term of the binomial, for any small $\epsilon > 0$, we have $u_{2n} \leq \frac{1+\epsilon}{2} \left(\frac{3}{\pi n}\right)^{3/2}$, eventually. It follows via the integral test that the sum $\sum_n u_{2n}$ is now convergent! This surprising result was proved by George Pólya in 1921.⁶

PÓLYA'S THEOREM *A random walk in one or two dimensions returns to the origin with probability one (and hence returns infinitely often to the origin with probability one). In three or more dimensions the probability that a walk returns to the origin is strictly less than one.*

For example, numerical calculations show only about one in three walks in three dimensions will return to the origin (the numerical probability is about 0.35).

10 Problems

1. With $0 < t < 1$ a fixed constant suppose the variable X has the following arithmetic distributions in turn: (a) $p(k) = \binom{n+k-1}{k} t^n (1-t)^k$ for $k \geq 0$; (b) $p(k) = 1/(k(k+1))$ for $k \geq 1$; (c) $p(k) = (1-t)t^{|k|}/(1+t)$ for integer k . Determine the generating function $\mathfrak{P}(s) = E(s^X)$ and thence the mean and variance in each case.

2. *Craps.* If two standard six-sided dice are thrown, the sum X of their face values has the arithmetic distribution Q given by Table 1 of Section I.2. Show that its generating function is $\mathfrak{Q}(s) = E(s^X) = s^2(1+s)^2(1+s+s^2)^2(1-s+s^2)^2/36$.

3. *Continuation, unusual dice.* Say that a die is *permissible* if it has six equally likely faces, each face showing a strictly positive integer which may be repeated on other faces. For example, a standard die is permissible and has faces $(1, 2, 3, 4, 5, 6)$; a die with faces $(1, 3, 3, 3, 5, 9)$ is non-standard but permissible. Demonstrate two *non-standard*, permissible dice (they will not be identical) whose sum of face values has the same distribution Q as that given in Table 1 of Section I.2. Show that your construction is the unique non-standard solution. [Hint: Determine an alternative *valid* factorisation of $\mathfrak{Q}(s)$.⁷]

4. A particle performs a random walk on the corners of a square ABCD at each step moving to one of the two neighbouring corners governed by probabilities attached to the edges. The probability is the same along any given edge for motion in either direction. Suppose (A,B) and (C,D) have a common edge probability p and (A, D) and (B, C) have a common edge probability $q = 1 - p$. Let u_n denote the probability that a particle starting at A returns to A in n steps. Show that the corresponding generating function satisfies

$$\mathfrak{U}(s) = \sum_n u_n s^n = \frac{1}{2} \left(\frac{1}{1-s^2} + \frac{1}{1-(q-p)^2 s^2} \right).$$

Hence determine the generating function of the time of first return to A.

⁶G. Pólya, "Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Strassennetz", *Mathematische Annalen*, vol. 84, pp. 149–160, 1921.

⁷Liu Tian told me of this curious problem and Jin You pointed me towards its elegant resolution.

5. Sums over a random number of terms. Let $S_N = X_1 + \dots + X_N$ where X_1, X_2, \dots is a sequence of positive, independent random variables drawn from a common distribution F , and N is a positive, arithmetic random variable of distribution G independent of the sequence $\{X_j, j \geq 1\}$. Suppose that F has mean α and variance σ^2 and G has mean m and variance s^2 . Determine $E(S_N)$ and $\text{Var}(S_N)$. The most important case arises when N is Poisson in which case S_N is said to have the *compound Poisson distribution*.

6. Continuation. Determine the d.f. $R(t) = P\{S_N \leq t\}$ in terms of F and G . Thence or otherwise determine $\widehat{R}(\zeta) = E(e^{-\zeta S_N})$ in terms of \widehat{F} and \widehat{G} .

7. Continuation. Specialise the results to the case where G is the (shifted) geometric distribution $P\{N = k\} = q^{k-1}p$ for $k \geq 1$ where, as usual, $0 < p < 1$ and $q = 1 - p$, and F is exponential with mean α . (a) Determine $\widehat{R}(\zeta)$ and thence determine $E(S_N)$ and $\text{Var}(S_N)$. (b) Show that S_N has an absolutely continuous distribution and determine its density. Verify the mean and variance by direct integrations of the density.

8. Suppose X_1, X_2, \dots are independent with common d.f. F , N is positive and arithmetic with generating function $\mathfrak{P}(s) = E(s^N)$. If N is independent of the X_i then $\max\{X_1, \dots, X_N\}$ has distribution $\mathfrak{P} \circ F$.

9. The branching process. In a stylised model of population growth we start with a single progenitor at generation 0 whose progeny form the first generation. Each of his progeny then can contribute 0, 1, 2, or more individuals to the second generation, the process continuing indefinitely. Call the number of progeny of any given individual his nuclear family. We suppose that the sizes of the nuclear families are independent random variables governed by the common arithmetic distribution $p(k)$. Let $1 = Z_0, Z_1, \dots, Z_n, \dots$ denote the successive generational sizes as the population evolves. Write $\mathfrak{G}_n(s) = E(s^{Z_n})$ for the generating function of the distribution of the size of the n th generation; here $\mathfrak{G}_1(s) = E(s^{Z_1}) = \sum_k p(k)s^k = \mathfrak{G}(s)$ represents the generating function of the distribution of nuclear family size. Show that $\mathfrak{G}_n(s) = \mathfrak{G}(\mathfrak{G}_{n-1}(s))$ and hence that $\mathfrak{G}_n(s) = \mathfrak{G}^{\circ n}(s)$ is obtained by the n -fold composition of \mathfrak{G} with itself.

10. Continuation, geometric branching. Suppose $p(k) = q^k p$ for $k \geq 0$ (as usual, $q = 1 - p$). Show by induction that

$$\mathfrak{G}_n(s) = \begin{cases} \frac{n-(n-1)s}{n+1-ns} & \text{if } p = q = 1/2, \\ \frac{q(p^n - q^n - ps(p^{n-1} - q^{n-1}))}{p^{n+1} - q^{n+1} - ps(p^n - q^n)} & \text{if } p \neq q. \end{cases}$$

Hence determine the probability of ultimate extinction $\lim_{n \rightarrow \infty} P\{Z_n = 0\}$ in each case.

11. The lognormal distribution is not determined by its moments. Suppose X is a positive variable with density $f(x) = (2\pi)^{-1/2}x^{-1}e^{-\log(x)^2/2}$ for $x > 0$. (This is the lognormal density; $\log X$ has the normal density.) Put $f_\alpha(x) = f(x)[1 + \alpha \sin(2\pi \log x)]$ with $-1 < \alpha < 1$. Show that f_α is a density with the same moments as f .

12. The renewal equation. Consider the distribution $U(t) = \sum_{k=0}^{\infty} F^{*k}(t)$ induced by a probability distribution F concentrated on the positive half-line. Working with the right-hand side of (3.2), show that if G is bounded then the convolution property of the Laplace transform implies that the renewal equation (3.1) has solution $V(t) = (U * G)(t)$.

13. Sojourn time. For the queuing process of Section 6, let $V_n = W_n + X_n$ be the sojourn time of the n th customer from the time of arrival to the time of departure. Determine an expression for the limiting Laplace transform of V_n as $n \rightarrow \infty$.

14. Continuation. Hence determine the expected sojourn time of a customer in the limit as $n \rightarrow \infty$ and verify the result by the Pollaczek–Khinchin formula.

15. Symmetric distributions. Suppose X and $-X$ have a common distribution F . Show that F is symmetric, that is, $F(x) = 1 - F(-x)$ at all points of continuity. If X is integrable argue thence that $E(X) = 0$.

16. Continuation. Suppose F and G are symmetric, not necessarily continuous. Show that $F * G$ is symmetric and hence that F^{*n} is symmetric for each n .

17. First ladder index. Suppose F is a continuous and symmetric distribution. Show that the first ladder index of the corresponding random walk has distribution $f(n) = \frac{1}{n} \binom{2n-2}{n-1} 2^{-2n+1}$ ($n \geq 1$). See Problems VIII.31,32.

18. Non-symmetric random walks. Suppose $P\{S_n > 0\} = \alpha$ is independent of n . Show that the generating function of the distribution of ladder indices satisfies $\mathfrak{U}(s) = (1-s)^{-\alpha}$ and hence that $P\{M_n = m\} = (-1)^n \binom{-\alpha}{m} \binom{-1+\alpha}{n-m}$.

19. Empirical distributions. In the notation of Problem XII.20 let $D_{n,n}$ denote the maximum discrepancy between two empirical distributions. Show that $P\{D_{n,n} \leq r/n\}$ is given by the probability that in a symmetric random walk in steps of ± 1 a path of length $2n$ starting and terminating at the origin does not reach the points $\pm \lfloor r \rfloor$. This is a famous theorem of B. V. Gnedenko and V. S. Koroljuk.⁸ [Hint: Order the $2n$ variables $X_1, \dots, X_n, X_1^#, \dots, X_n^#$ in order of increasing magnitude and put the ordered sequence in one-to-one correspondence with a walk where the j th step is $+1$ if the corresponding variable in the ordered sequence appears unhashed and -1 if the variable is hashed.]

Problems 20–27 deal with the theory of runs. A convention permits the seamless application of the theory of renewals to this setting. In a succession of Bernoulli trials we say that a first success run of length r occurs at trial n if a succession of r successes occurs for the first time at trials $n-r+1, n-r+2, \dots, n$. The occurrence of a first success run at epoch n triggers a renewal and we consider a statistically identical process to restart with trial $n+1$ and, thereafter, each time a success run of length r is first observed. For instance, the renewal epochs for success runs of length 3 are identified by a comma followed by a little added space for visual delineation in the following sequence of Bernoulli trials: 0110111, 1000010111, 0111, 111, 101 ···. With this convention, success runs of length r determine a renewal process. A similar convention and terminology applies to failure runs, as well as to runs of either type.

In the following problems we consider a sequence of Bernoulli trials with success probability p and failure probability $q = 1 - p$. Suppose $r \geq 1$. For each n , write $u_{r,n}^{(0)}$, $u_{r,n}^{(1)}$, and $u_{r,n}$ for the probabilities that there is a failure run, success run, or run of either type, respectively, of length r at trial n . Let $\mathfrak{U}_r^{(0)}(s)$, $\mathfrak{U}_r^{(1)}(s)$, and $\mathfrak{U}_r(s)$ denote the corresponding generating functions. Likewise, let $f_{r,n}^{(0)}$, $f_{r,n}^{(1)}$, and $f_{r,n}$ represent the probabilities that there is a first failure run, success run, or run of either type, respectively, at trial n , with $\mathfrak{F}_r^{(0)}(s)$, $\mathfrak{F}_r^{(1)}(s)$, and $\mathfrak{F}_r(s)$ the corresponding generating functions. Finally, let $q_{r,n}^{(0)}$, $q_{r,n}^{(1)}$, and $q_{r,n}$ denote the probabilities of no failure run, success run, or run of either type, respectively, through trial n .

⁸B. V. Gnedenko and V. S. Koroljuk, “On the maximum discrepancy between two empirical distributions”, *Selected Translations in Mathematical Statistics and Probability*, IMS and AMS, vol. 1, pp. 13–16, 1961.

20. *Success runs.* Argue that $p^r = u_{r,n}^{(1)} + u_{r,n-1}^{(1)}p + \dots + u_{r,n-r+1}^{(1)}p^{r-1}$ for $n \geq r$. Hence show that $(1-s)(1-p^rs^r)\mathfrak{U}_r^{(1)}(s) = 1-s+qp^rs^{r+1}$.

21. *Continuation.* Show that $\mathfrak{F}_r^{(1)}(s) = 1 - 1/\mathfrak{U}_r^{(1)}(s)$, hence $\mathfrak{F}_r^{(1)}(s) = P_r(s)/Q_r(s)$ is a rational form with $P_r(s) = p^rs^r$ and $Q_r(s) = 1 - qs(1 + ps + \dots + p^{r-1}s^{r-1})$.

22. *Continuation.* Determine the mean and variance of the recurrence times of success runs of length r .

23. *Continuation, asymptotics.* Show that $Q_r(s)$ has a unique positive root x and that this root is smaller in absolute value than any other (complex) root. By an expansion of $Q_r(s)^{-1}$ in partial fractions conclude that the asymptotic behaviour of $f_{r,n}^{(1)}$ is determined by x and show that $f_{r,n}^{(1)} \sim \frac{-p^rx^r}{Q'_r(x)} \cdot \frac{1}{x^{n+1}}$. Using the fact that $Q_r(x) = 0$ show thence that $f_{r,n}^{(1)} \sim \frac{(x-1)(1-px)}{q(r+1-rx)} \cdot \frac{1}{x^{n+1}}$ as $n \rightarrow \infty$.

24. *Continuation, run-free sequences.* Show that $q_{r,n}^{(1)} \sim \frac{1-px}{q(r+1-rx)} \cdot \frac{1}{x^{n+1}}$ as $n \rightarrow \infty$. If $p = 1/2$ conclude that there is a less than one in five chance of not encountering at least one success run of length five in 100 trials. Now revisit Problems I.17,18.

25. *Runs of either type.* Argue that $\mathfrak{U}_r(s) = \mathfrak{U}_r^{(0)}(s) + \mathfrak{U}_r^{(1)}(s) - 1$ and hence determine an explicit rational form for $\mathfrak{F}_r(s)$.

26. *Continuation.* Determine the mean recurrence time for a run of either type of length r .

27. *Continuation, the symmetric case.* If $p = 1/2$ show that $\mathfrak{F}_r(s) = s \cdot \mathfrak{F}_r^{(1)}(s)$ and hence that $f_{r,n} \sim \frac{(x_{r-1}-1)(2-x_{r-1})}{r-(r-1)x_{r-1}} \cdot \frac{1}{x_{r-1}^n}$ and $q_{r,n} \sim \frac{2-x_{r-1}}{r-(r-1)x_{r-1}} \cdot \frac{1}{x_{r-1}^n}$ where, for each k , x_k denotes the smallest positive root of the polynomial $Q_k(s) = 1 - \frac{s}{2}[1 + \frac{s}{2} + \dots + (\frac{s}{2})^{k-1}]$. By numerical evaluation conclude that the chance that there is no run of either type of length five in 100 trials is less than 3%. This illustrates once more the surprising character of fluctuations; most readers feel that such long runs are "unnatural".

Problems 28–31 deal with the characteristic function of a distribution F . Suppose X has (arbitrary) distribution F . Reusing notation, we define the Fourier–Lebesgue transform of the distribution by $\widehat{F}(\xi) = E(e^{i\xi X}) = \int_{-\infty}^{\infty} e^{i\xi x} dF(x)$ where, as usual, $i = \sqrt{-1}$ is the imaginary root of unity. (If F has a density f then \widehat{F} is just the ordinary Fourier transform \widehat{f} of f .) We call \widehat{F} the characteristic function of X or the associated distribution F .

28. *Characteristic functions.* Suppose the d.f. F has characteristic function \widehat{F} . Prove the following properties: (a) $\widehat{F}(0) = 1$ and $|\widehat{F}(\xi)| \leq 1$ for all ξ ; (b) \widehat{F} is uniformly continuous on \mathbb{R} ; (c) for all real ξ_1, \dots, ξ_n and complex z_1, \dots, z_n , we have $\sum_{j,k} \widehat{F}(\xi_j - \xi_k) z_j z_k^* \geq 0$, in other words, the characteristic function is positive definite.

29. *Independence.* If X and Y are independent with distributions F and G , respectively, then $X + Y$ has characteristic function $\widehat{F} \cdot \widehat{G}$. Is the converse statement true?

30. *Affine shifts.* Suppose X has characteristic function $\widehat{F}(\xi)$. Then $aX + b$ has characteristic function $e^{i\xi b} \widehat{F}(a\xi)$.

31. *Differentiation.* Suppose X has characteristic function $\widehat{F}(\xi)$. If X has an n th moment then $E(X^n) = i^n \widehat{F}^{(n)}(0)$.

XVI

The Law of Large Numbers

The link between axiomatic probability and the frequentist view of probability is provided by the law of large numbers that we first encountered in Chapter V. The basic limit laws are not only important in their own right but, as we shall see in this chapter, inform a whole slew of applications. The simple inequality of Chebyshev is key to these limit laws and we begin by revisiting it.

C 1, 2, 9

A 3–7

§ 8, 10–12

1 Chebyshev's inequality, reprise

Suppose f and g are Baire functions, f dominated by g . If $f(X)$ and $g(X)$ are integrable random variables then monotonicity of expectation says that $E(f(X)) \leq E(g(X))$. This humble inequality lays the foundation for deep tail estimates.

THEOREM 1 *Suppose $g(x)$ is any Baire function dominating the shifted Heaviside function $H_0(x - 1)$. Then $P\{X \geq t\} \leq E g(\frac{X}{t})$ for any $t > 0$.*

PROOF: As $H_0(tx) = H_0(x)$ for every $t > 0$, we have $H_0(x-t) = H_0[t(\frac{x}{t}-1)] = H_0(\frac{x}{t}-1) \leq g(\frac{x}{t})$. As $P\{X \geq t\} = E H_0(X-t)$, the stated conclusion follows by monotonicity of expectation. ►

The proof is so simple that it seems almost silly to dignify the result as a theorem but its elementary nature belies its impact. It is natural to try to obtain a feel for the result by trying out various natural choices for the dominating function g . The simplest choice is a linear ramp.

MARKOV'S INEQUALITY *Suppose X is a positive random variable, $t > 0$ a positive constant. Then $P\{X \geq t\} \leq E(X)/t$.*

PROOF: The linear ramp $g(x) = x$ for $x \geq 0$ and $g(x) = 0$ for $x < 0$ dominates $H_0(x-1)$. Consequently,

$$P\{X \geq t\} \leq \int_{\mathbb{R}^+} \frac{x}{t} dF(x) = \frac{E(X)}{t},$$

the conclusion in the final step following as X is concentrated on \mathbb{R}^+ . ▶

Markov's inequality is flexible but not very strong. If we couple it with exponential transformations, however, we can get results of quite remarkable delicacy. I shall illustrate the idea here in the context of the binomial, both by way of preparing the ground for a more detailed study in Section XVII.1, as well as in view of the basic importance of the result in its own right.

Suppose X_1, \dots, X_n are symmetric Bernoulli trials corresponding to fair coin tosses, their sum S_n representing the number of accumulated successes. The binomially distributed variable S_n has mean $n/2$ and it is of interest to estimate the probability that S_n deviates from its mean by nt or more for some $t \geq 0$. This is the province of large deviations. Fix any $\lambda \geq 0$. We may now write down the string of inequalities

$$\begin{aligned} \mathbf{P}\{S_n \geq n/2 + nt\} &\stackrel{(i)}{=} \mathbf{P}\{e^{\lambda(S_n - n/2 - nt)} \geq 1\} \stackrel{(ii)}{\leq} e^{-\lambda nt} \mathbf{E}\left(\prod_{j=1}^n \exp(\lambda(X_j - 1/2))\right) \\ &\stackrel{(iii)}{=} e^{-\lambda nt} \prod_{j=1}^n \mathbf{E}(e^{\lambda(X_j - 1/2)}) = e^{-\lambda nt} \left(\frac{e^{\lambda/2} + e^{-\lambda/2}}{2}\right)^n \stackrel{(iv)}{\leq} e^{-\lambda nt + \lambda^2 n/8} \end{aligned}$$

whose justifications follow: (i) by exponentiating both sides as monotone transformations preserve inequalities; (ii) by Markov's inequality applied to the positive variable $e^{-\lambda nt} \cdot e^{\lambda(S_n - n/2)}$; (iii) as functions of independent variables are independent; and (iv) via the simple observation $(e^{\lambda/2} + e^{-\lambda/2})/2 \leq e^{\lambda^2/8}$, as may be verified, for example, by comparing coefficients in the Taylor expansions of the two sides. As the inequality holds for all $\lambda \geq 0$, we may take the infimum of both sides to obtain the elegantly simple bound

$$\mathbf{P}\{S_n \geq n/2 + nt\} \leq \inf_{\lambda \geq 0} e^{-\lambda nt + \lambda^2 n/8} = e^{-2nt^2}$$

by a straightforward minimisation. An identical bound for the left tail $\mathbf{P}\{S_n \leq n/2 - nt\}$ holds by symmetry.

HOEFFDING'S INEQUALITY FOR THE BINOMIAL Let S_n denote the accumulated number of successes in n tosses of a fair coin. Then, for every $t \geq 0$, we have

$$\mathbf{P}\{|S_n - n/2| \geq nt\} \leq 2e^{-2nt^2}. \quad (1.1)$$

The simple demonstration given here was noted by R. M. Dudley.

The exponential bound for the binomial tail implies a corresponding result for large deviations of a simple random walk from the origin. Suppose Z_1, Z_2, \dots is a sequence of independent random variables drawn from the common distribution placing equal mass at the points -1 and $+1$. The partial sum $R_n = Z_1 + \dots + Z_n$ then represents a simple random walk over n steps which is related to the symmetric binomial via $R_n = 2S_n - n$. In view of Hoeffding's inequality for the binomial we see that

$$\mathbf{P}\{|R_n| \geq \xi\} = \mathbf{P}\{|S_n - n/2| \geq \xi/2\} \leq 2e^{-\xi^2/(2n)}. \quad (1.1')$$

The only place Hoeffding's inequality is needed in this chapter is in Section 11 but the reader will find it in repeated use in the next chapter.

Estimates of deviations from the mean instead of tail probabilities may be obtained by the simple expedient of replacing X in Theorem 1 by $|X - E(X)|$. The choice of quadratic as a dominating function gives an exceptionally versatile bound.

CHEBYSHEV'S INEQUALITY Suppose X has a finite second moment and t is strictly positive. Then $P\{|X - E(X)| \geq t\} \leq \text{Var}(X)/t^2$.

PROOF: The quadratic $g(x) = x^2$ clearly dominates $H_0(x - 1)$. If we make the replacement $|X - E(X)| \leftarrow X$ in Theorem 1 the conclusion is trite. ►

The reader has seen Chebyshev's idea in use already in Sections V.6, VIII.3, XIV.9, and XV.1. Theorem 1 merely puts the principle in sharp relief.

The inequality of Chebyshev provides support for the intuition that the standard deviation is a measure of the likely spread of X around its mean. Writing $\sigma^2 = \text{Var}(X)$, if we set $t = c\sigma$ then $P\{|X - E(X)| \geq c\sigma\} \leq 1/c^2$. The chances that X lies more than three standard deviations from its mean are no more than $1/9$ or about 10%.

Chebyshev proved his inequality in 1852¹ and few inequalities have aged so gracefully. Its virtue is not that it is sharp—we will see much tighter bounds in the sequel—but quite extraordinarily versatile in divers domains, illustrating once more the moral that simple bounds are frequently much more useful than condition-bound, ponderous results. The following selection of applications of Chebyshev's inequality is somewhat more sophisticated than those seen hitherto and may help sway the sceptical reader who is not yet convinced of its general utility.

2 Khinchin's law of large numbers

Chebyshev's inequality is particularly effective when we deal with sums of independent (or even just uncorrelated) variables. Suppose X_1, X_2, \dots are independent random variables with a common distribution F . As usual, let $S_n = X_1 + \dots + X_n$ for each n . If the variables X_k are square-integrable with mean μ and variance σ^2 then $\frac{1}{n}S_n$ has mean μ and variance σ^2/n . Chebyshev's inequality now lets us conclude that

$$P\left\{\left|\frac{1}{n}S_n - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}. \quad (2.1)$$

The bound on the right tends to zero as $n \rightarrow \infty$ for every choice of $\epsilon > 0$ and generalising the result of Section V.6 we conclude that $\frac{1}{n}S_n$ converges in probability to μ . A. Ya. Khinchin proved in 1929 that the requirement that the

¹For an account from its discoverer see P. L. Chebyshev, "Des valeurs moyennes", *Journal de Mathématiques Pures et Appliquées*, vol. 12, pp. 177–184, 1867.

variables X_k have variance can be relaxed and the result now takes on a very elegant and simple form.²

THE WEAK LAW OF LARGE NUMBERS Suppose X_1, X_2, \dots is a sequence of independent random variables drawn from a common distribution with finite mean μ and let $S_n = X_1 + \dots + X_n$ for each n . Then $\frac{1}{n}S_n \rightarrow^p \mu$ or, spelled out, $\mathbf{P}\left\{\left|\frac{1}{n}S_n - \mu\right| \geq \epsilon\right\} \rightarrow 0$ as $n \rightarrow \infty$ for every choice of $\epsilon > 0$.

Khinchin's result was very soon to be superseded but here, as elsewhere, the first step is the critical one. While I will present Kolmogorov's strengthening of the law of large numbers later in this chapter, it is yet worthwhile to see Khinchin's proof without the main ideas obscured by technical complications. The proof is not merely of historical interest; the method continues to find frequent use in the proof of limit theorems.

Chebyshev's inequality cannot be usefully applied to the sum of the X_j if the variables don't have finite variance. The versatile *method of truncation* finesse this by creating auxiliary variables that do. Let $\{a_n, n \geq 1\}$ be a divergent positive sequence to be determined. For each j we set

$$U_j = \begin{cases} X_j & \text{if } |X_j| \leq a_n, \\ 0 & \text{if } |X_j| > a_n, \end{cases} \quad \text{and} \quad V_j = \begin{cases} 0 & \text{if } |X_j| \leq a_n, \\ X_j & \text{if } |X_j| > a_n, \end{cases} \quad (2.2)$$

or, more compactly, defining the event $A_n = \{|X| \leq a_n\}$, we have $U_j = X_j 1_{A_n}(X_j)$ and $V_j = X_j 1_{A_n^c}(X_j)$, whence $X_j = U_j + V_j$. Now

$$|X_1 + \dots + X_n - \mu n| \leq |U_1 + \dots + U_n - \mu n| + |V_1 + \dots + V_n|$$

by the triangle inequality and by Boole's inequality it follows that

$$\begin{aligned} \mathbf{P}\left\{\left|\frac{1}{n}S_n - \mu\right| \geq \epsilon\right\} &\leq \mathbf{P}\left\{|U_1 + \dots + U_n - \mu n| \geq \frac{1}{2}\epsilon n\right\} \\ &\quad + \mathbf{P}\left\{|V_1 + \dots + V_n| \geq \frac{1}{2}\epsilon n\right\} \end{aligned} \quad (2.3)$$

as the occurrence of the event on the left implies the occurrence of at least one of the events on the right. We leverage the fact that the variables X_j are integrable to estimate the two terms on the right in turn.

As $M = \int_{\mathbb{R}} |x| dF(x)$ is finite the integral tails must decay to zero. Thus, for any $\zeta > 0$, we must have $\int_{|x|>a_n} |x| dF(x) < \zeta$ as $n \rightarrow \infty$. It follows that

$$\mathbf{P}\{|X_j| > a_n\} = \int_{|x|>a_n} dF(x) \leq \int_{|x|>a_n} \frac{|x|}{a_n} dF(x) < \frac{\zeta}{a_n},$$

eventually, for all sufficiently large n .

²A. Khinchin, "Sur la loi des grands nombres", *Comptes rendus de l'Académie des Sciences*, vol. 189, pp. 477–479, 1929.

We begin with the second of the terms on the right in (2.3). As the occurrence of the event $|V_1 + \dots + V_n| \geq \frac{1}{2}\epsilon n$ certainly implies that at least one of the V_j is non-zero, Boole's inequality implies

$$\mathbf{P}\{|V_1 + \dots + V_n| \geq \frac{1}{2}\epsilon n\} \leq \mathbf{P}\left(\bigcup_{j=1}^n \{V_j \neq 0\}\right) \leq \sum_{j=1}^n \mathbf{P}\{|X_j| > a_n\} < \frac{\zeta n}{a_n}.$$

For the bound to be useful we conclude that a_n must have a growth rate at least of the order of n .

Now each U_j has support in a bounded interval and *a fortiori* has a finite second moment. Chebyshev's inequality can hence be deployed to estimate the first term on the right in (2.3). If we set $\mu_n = \mathbf{E}(U_j) = \int_{|x| \leq a_n} x dF(x)$ then

$$|\mu_n - \mu| = \left| \int_{|x| > a_n} x dF(x) \right| \leq \int_{|x| > a_n} |x| dF(x) < \zeta \quad (\text{eventually}).$$

We can afford to be more cavalier in estimating the variance. As $x^2 \leq a_n|x|$ if $|x| \leq a_n$, by monotonicity of expectation we have

$$\begin{aligned} \text{Var}(U_j) &\leq \mathbf{E}(U_j^2) = \int_{|x| \leq a_n} x^2 dF(x) \\ &\leq a_n \int_{|x| \leq a_n} |x| dF(x) \leq a_n \int_{-\infty}^{\infty} |x| dF(x) = Ma_n. \end{aligned}$$

The pieces are in place for an application of Chebyshev's inequality. We first centre the sum appropriately by a routine application of the triangle inequality,

$$|U_1 + \dots + U_n - \mu n| \leq |U_1 + \dots + U_n - \mu_n n| + |\mu_n n - \mu n|.$$

For any $0 < \zeta < \epsilon/4$, Chebyshev's inequality now yields

$$\begin{aligned} \mathbf{P}\{|U_1 + \dots + U_n - \mu n| \geq \frac{1}{2}\epsilon n\} &\leq \mathbf{P}\{|U_1 + \dots + U_n - \mu_n n| \geq \frac{1}{2}\epsilon n - |\mu_n - \mu|n\} \\ &\leq \mathbf{P}\{|U_1 + \dots + U_n - \mu_n n| \geq \frac{1}{2}(\epsilon - 2\zeta)n\} \leq \frac{4Mna_n}{(\epsilon - 2\zeta)^2 n^2} \leq \frac{16Ma_n}{\epsilon^2 n}. \end{aligned}$$

The bound will diverge if a_n has a growth rate faster than n and so a_n cannot increase faster than the order of n . From our two bounds we conclude that, if the approach is to be useful at all, a_n must be chosen to be exactly of order n . Accordingly, set $a_n = tn$ with t to be specified. Pooling our bounds, we obtain

$$\mathbf{P}\left\{ \left| \frac{1}{n} S_n - \mu \right| \geq \epsilon \right\} < \frac{16Ma_n}{\epsilon^2 n} + \frac{\zeta n}{a_n} = \frac{16Mt}{\epsilon^2} + \frac{\zeta}{t}.$$

The parameters t and ζ may be chosen at our discretion and we select them to make the right-hand side small. Fix any tiny, strictly positive δ . If we select

$t = \epsilon^2 \delta / 32M$ then the first term on the right is no larger than $\delta/2$ and if we set ζ to be the smaller of $\epsilon/4$ and $\delta t / 2 = \epsilon^2 \delta^2 / 64M$ then so is the second term on the right. It follows that $P\left\{ \left| \frac{1}{n} S_n - \mu \right| \geq \epsilon \right\} < \delta$, eventually, for all sufficiently large n . As δ may be chosen arbitrarily small, the proof is concluded.

The law of large numbers validates the vague but intuitive idea that “superposing independent observations averages out the noise”. In many ways this is the most intuitive portion of the theory aligning as it does the axiomatic approach to probability with the intuitive frequentist approach: the “sample mean” $\frac{1}{n}(X_1 + \dots + X_n)$ is indeed a good probabilistic estimate of the “ensemble mean” μ if n is suitably large. The next section outlines an unexpected—and very useful—application of this idea.

3 A physicist draws inspiration from Monte Carlo

There is perhaps no other domain of endeavour which requires the computation of integrals in such a large number of dimensions as quantum physics. Suppose $f(x) = f(x_1, \dots, x_v)$ is a bounded, integrable function on the unit cube $[0, 1]^v$ in v dimensions where, by a suitable scaling, we may assume $|f(x)| \leq 1$ everywhere in the unit cube. In a typical setting one wishes to evaluate an integral of the form

$$J = \int \cdots \int_{[0, 1]^v} f(x_1, \dots, x_v) dx_1 \cdots dx_v. \quad (3.1)$$

Computations of this nature are important in nuclear engineering and took on a particular importance after the Second World War. How should one proceed?

An explicit closed-form solution is too much to hope for in general and so we are driven to numerical estimates. To begin, let us suppose that f is explicitly known or, at least, computable at any point $x = (x_1, \dots, x_v)$ with only a modest amount of effort. Experience with mathematical software packages now suggests an iterated numerical evaluation of the integrals. It does not matter much which procedure we pick, but using, say, Simpson’s rule will require the evaluation of f at three points for each integral so that the procedure will require 3^v evaluations of f in total. While the number of computations is certainly small when $v = 1$ or 2 , the computational demands rapidly get out of control as v increases: for $v = 20$ the procedure already requires more than three billion evaluations of f while for $v = 200$ the number of evaluations needed is a mind-boggling 2.6×10^{95} . As some estimates give the number of atoms in the observable universe to be about 10^{80} such a number is not only beyond the reach of our current computing resources but may be forever beyond reach. The picture does not change if the reader replaces Simpson’s rule by the trapezoidal rule or some other procedure from her elementary calculus class. While these procedures vary in detail they all require one to evaluate f on a grid and the number of grid points grows ineluctably exponentially in dimension.

The computational intractability of the classical procedures in this setting does not eliminate the need. The aftermath of the Second World War saw a great surge in nuclear research and the need for computations of the form (3.1) was exigent. It was in this climate that Enrico Fermi, Stanislaw Ulam, and John von Neumann came up with the bewildering idea of a chance-driven computation.

Suppose $\mathbf{X} = (X_1, \dots, X_v)$ is a random point drawn by sampling uniformly from the unit cube $[0, 1]^v$. Then $f(\mathbf{X})$ is a bounded random variable whose expected value $E(f(\mathbf{X}))$ is exactly J and whose variance is bounded by $\text{Var}(f(\mathbf{X})) \leq E(f(\mathbf{X})^2) \leq 1$ by monotonicity of expectation as $|f|$ is bounded by 1. This suggests a pathway to approximating the value of the integral. Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(n)}$ be a sequence of random points selected by independent sampling from the uniform distribution in $[0, 1]^v$. The sample mean of the function values at these points is now a tempting surrogate for J ; indeed, with $S_n = f(\mathbf{X}^{(1)}) + \dots + f(\mathbf{X}^{(n)})$, the weak law estimate (2.1) shows that

$$P\left\{\left|\frac{1}{n}(f(\mathbf{X}^{(1)}) + \dots + f(\mathbf{X}^{(n)})) - J\right| \geq \epsilon\right\} \leq \epsilon^{-2} n^{-1}$$

and, by selecting n suitably large, the right-hand side can be made as small as desired. Thus, by sampling the values of f at $n = 200,000$ random points in the cube one can obtain an estimate $\frac{1}{n}S_n$ that approximates J with an error ϵ of no more than 1% and a confidence $1 - \delta = 1 - \epsilon^{-2}n^{-1}$ of at least 95%. In $v = 200$ dimensions this translates into the generation of $200 \times 200,000 = 4 \times 10^7$ uniform variates and 200,000 evaluations of f —numbers which, while they would have stretched the capabilities of a computer circa 1950, are well within the computational reach of even the skimpiest modern computer.

The fact that there is a tiny error in the approximation is *not* what is germane—any finite numerical procedure is bound to lead to an approximate result. What is so different about the approach is that it is fundamentally *probabilistic*—there is a possibility, albeit small, that the estimate of J may be completely inaccurate. This certainly introduces a new wrinkle: while hitherto any computational procedure connoted both precision and certainty, this procedure was both approximate and uncertain. This is a little uncomfortable to be sure but we may again take comfort in the bromide that a principled answer in a small number of steps is preferable to no answer at all in any number of steps. Von Neumann christened the procedure the *Monte Carlo method* in view of its chance-driven overtones; the name has now come to mean a class of chance-driven procedures of this type.

The Monte Carlo method exhibits the remarkable feature that *the requisite sample size is independent of dimension*; the number of evaluations is the same whether we deal with two dimensions or two hundred ... or two million! This is counterbalanced somewhat by the observation that, for a given level of confidence $1 - \delta$, the error ϵ decreases only as the reciprocal of the square root of the sample size n . Thus, to get an additional decimal place of accuracy we will

require a hundred-fold increase in the sample size.

SLOGAN *The Monte Carlo method provides good computational estimates quickly, with little effort, but it is not possible to get additional precision through the method without huge cost.*

Monte Carlo methods have now become a staple in a variety of fields from quantum physics to the large-scale modelling of galaxies, from the analysis of integrated circuits to the modelling of fluid flows, in the qualitative analysis of risk, and in the development of randomised sampling methods for auditing large transactional records. With the rise in importance of probabilistic computing arose the need to be able to quickly and accurately generate synthetic random numbers. It then became important to answer the fundamental question: “What is a random number?” The search for a proper definition and the rise of an attendant theory sparked the development of modern pseudorandom number generators. Principal among these are the *linear congruential generators* which, starting from an arbitrary initial “seed” X_0 , sequentially generate a sequence of numbers by a fixed recurrence of the type $X_{n+1} = (aX_n + c) \bmod m$. The proper selection of the multiplier a , the increment c , and the modulus m so that the resulting deterministic sequence $\{X_n, n \geq 1\}$ appears truly random is an art form guided by some theory.³ Random number generators of this type are now routinely embedded in the pervasive computing devices of today.

4 Triangles and cliques in random graphs

Consider the Erdős–Rényi random graph $G_{n,p}$ of Section IV.7. Say that a collection of three vertices $S = \{i, j, k\}$ forms a *triangle* if the graph $G_{n,p}$ contains all three edges $\{i, j\}$, $\{i, k\}$, and $\{j, k\}$. What is the probability that the graph is triangle-free, that is, contains no triangle? Intuitively, if p is large then the graph is quite likely to contain a triangle; if p is small, however, it is likely to be triangle-free. The question is what is the critical range for $p = p_n$ at which triangles first begin to appear?

Let S be any set of three vertices, X_S the indicator for the event that S forms a triangle. Then X_S represents a Bernoulli trial with success probability p^3 as S forms a triangle if, and only if, all three edges connecting the vertices in S are present in the graph. In particular, $E(X_S) = p^3$ and $\text{Var}(X_S) \leq E(X_S^2) = p^3$.

Now let N be the number of triangles in $G_{n,p}$. Then $N = \sum_{S: \text{card}(S)=3} X_S$. As N is a positive arithmetic variable, by summing over sets S of three vertices, the probability that $G_{n,p}$ has one or more triangles is bounded by

$$P\{N \geq 1\} \leq E(N) = \sum_S E(X_S) = \binom{n}{3} p^3 < n^3 p^3. \quad (4.1)$$

³The interested reader will find a comprehensive account in D. Knuth, *The Art of Computer Programming: Volume 2, Seminumerical Algorithms*. Reading, MA: Addison-Wesley, 1981.

Accordingly, if $p = p_n \ll 1/n$ then $\mathbf{P}\{N \geq 1\} \ll 1$ and with high probability there are no triangles in the graph.

It is tempting to conjecture that if, on the other hand, $p = p_n \gg 1/n$ then (many) triangles appear but the basic expectation inequality in (4.1) doesn't go the other way. No matter, Chebyshev's inequality carries the argument through with panache. The probability that there are no triangles in the graph is given by

$$\mathbf{P}\{N = 0\} \leq \mathbf{P}\{|N - \mathbf{E}(N)| \geq \mathbf{E}(N)\} \leq \frac{\text{Var}(N)}{\mathbf{E}(N)^2} \quad (4.2)$$

and we wish to determine $p = p_n$ for which the right-hand side is asymptotically small.

By linearity of expectation, we may write

$$\text{Var}(N) = \sum_S \text{Var}(X_S) + \sum_S \sum_{T \neq S} \text{Cov}(X_S, X_T). \quad (4.3)$$

The first sum on the right is no larger than $\binom{n}{3}p^3 \leq n^3p^3$ and it only remains to estimate the covariances $\text{Cov}(X_S, X_T)$.

Suppose S and T are two distinct sets of three vertices. As shown in Figure 1, there are three possibilities depending on how many vertices S and T share in common. If S and T are disjoint or if S and T share a single vertex

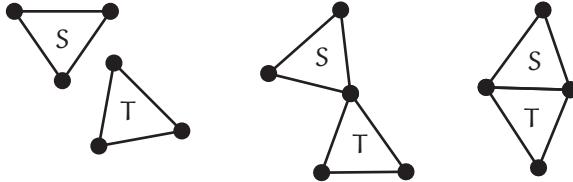


Figure 1: Triangles in a graph.

then the event that S forms a triangle is independent of the event that T forms a triangle as the triangle on S shares no edges with the triangle on T . Indeed, in these cases triangles are formed on both S and T if, and only if, each of the six distinct edges in play is in the graph and this has probability $p^6 = p^3 \cdot p^3$. Thus, if $\text{card } S \cap T \leq 1$ then the indicators X_S and X_T are independent, *a fortiori* uncorrelated, whence $\text{Cov}(X_S, X_T) = 0$. The entire contribution to the double sum on the right in (4.3) hence comes from the remaining case where S and T share exactly two vertices. But then there are precisely five edges in play and as each of these has to be present in the graph the probability that both S and T form triangles is given by p^5 ; the indicators X_S and X_T are now dependent. As the product $X_S X_T$ is simply the indicator for the joint occurrence of the event that S forms a triangle and T forms a triangle, we have $\text{Cov}(X_S, X_T) \leq \mathbf{E}(X_S X_T) = p^5$. We may select S in $\binom{n}{3} < n^3/3$ ways and, for each S , there are exactly $\binom{3}{2} \binom{n-3}{1} < 3n$

choices of T sharing exactly 2 vertices with S . There are hence no more than $\frac{1}{3}n^3 \cdot 3n = n^4$ non-zero contributing terms in the double sum and, consequently,

$$\sum_S \sum_{T \neq S} \text{Cov}(X_S, X_T) < n^4 p^5.$$

We need a lower bound on the denominator on the right in (4.2) to keep the inequality going in the right direction. Equally simple bounds suffice: we have

$$E(N) = \binom{n}{3} p^3 \geq \frac{1}{6}(n-3)^3 p^3 = \frac{1}{6}n^3 p^3 (1 - 3/n)^3 \geq \frac{1}{10}n^3 p^3$$

if $n \geq 20$. It follows that

$$P\{N = 0\} \leq \frac{100(n^3 p^3 + n^4 p^5)}{n^6 p^6} = \frac{100}{n^3 p^3} + \frac{1}{n} \cdot \frac{100}{np}$$

(at least for $n \geq 20$). And thus, if $p = p_n \gg 1/n$ then $P\{N = 0\} \ll 1$ and the graph exhibits triangles with high probability.

THEOREM Suppose $\{p_n\}$ is a sequence of probabilities. Then, as $n \rightarrow \infty$,

$$P\{G_{n,p_n} \text{ is triangle-free}\} \rightarrow \begin{cases} 1 & \text{if } \frac{p_n}{1/n} \rightarrow 0, \\ 0 & \text{if } \frac{p_n}{1/n} \rightarrow \infty. \end{cases}$$

Quite remarkably, a sharp *threshold phenomenon* (or, in the physicist's language, a *phase transition*) has emerged. If the edge probability $p = p_n$ decays with n slightly faster than the order of $1/n$ then, asymptotically for large n , almost all random graphs will be triangle-free; contrariwise, if $p = p_n$ decays with n slightly slower than the order of $1/n$ then, asymptotically, almost all random graphs will exhibit (many) triangles. This kind of sharp asymptotic threshold phenomenon is surprising at first sight but is quite typical in random graphs; the underlying reason is a deep structural property of high-dimensional spaces.

The reader may feel on consideration that rather more should be provable. For $p_n \sim c/n$ for any constant c the event that a triangle is formed on a given set of three vertices is quite small—this is the province of rare events, our covariance considerations showing that there are a few scattered dependencies though most of these events are pairwise independent. This is very reminiscent of the Poisson paradigm and it is tempting to conjecture that *the number of triangles in $G_{n,p}$ is governed asymptotically by the Poisson distribution with parameter $c^3/6$* . This is true. The reader who wishes to prove it using the methods of Chapter IV will need to estimate quantities of the form $\sum_{S_1, \dots, S_k} E(X_{S_1} \cdots X_{S_k})$ where the sum is over distinct three-vertex sets. If she does she will emerge with a deeper appreciation of how much labour Chebyshev's inequality has saved (at the cost of slightly less precision).

We could replace the event that there is a triangle by the event that there is a clique on four vertices, that is to say, a complete subgraph on four vertices or a K_4 , in the random graph $G_{n,p}$. And more generally, we could consider the event that $G_{n,p}$ contains a complete subgraph on k vertices, a K_k . The calculations, if slightly more involved, do not materially change. I will leave these to the *Problems* at the end of this chapter.

5 A gem of Weierstrass

Suppose f is infinitely differentiable in some neighbourhood of the origin. Then the formal power series $\sum_{k=0}^{\infty} f^{(k)}(0)x^k/k!$ is the Taylor series about the origin generated by f . The reader will no doubt recall from her elementary calculus classes that if f is sufficiently well-behaved then its Taylor series converges pointwise to f in a neighbourhood of the origin. The series, while undeniably useful, has significant limitations. First and foremost, it requires that f be extremely well-behaved and in the process rules out a very large number of useful functions. It is even more disquieting that the condition that f be infinitely differentiable does not in itself guarantee that the series will converge to f .

EXAMPLE: A smooth function with a misleading Taylor series. Let f be the function defined by $f(x) = e^{-1/x^2}$ for $x \neq 0$ and with $f(0) = 0$. L'Hôpital's rule quickly shows that f is infinitely differentiable. Indeed, routine applications of the chain rule show that

$$f'(x) = \frac{2}{x^3} e^{-1/x^2}, \quad f''(x) = \left(-\frac{6}{x^4} + \frac{4}{x^6}\right) e^{-1/x^2}, \quad f'''(x) = \left(\frac{24}{x^5} - \frac{36}{x^7} + \frac{8}{x^9}\right) e^{-1/x^2}.$$

By induction it is easy to verify that $f^{(k)}(x) = P_k(x^{-1})e^{-1/x^2}$ where P_k is a polynomial of degree $3k$. As $x^{-m}e^{-1/x^2} \rightarrow 0$ as $x \rightarrow 0$ for each fixed m , it follows that $f^{(k)}(0) = 0$ for every $k \geq 1$. In consequence, the Taylor series $\sum_{k=0}^{\infty} f^{(k)}(0)x^k/k!$ is identically zero for all x . On the other hand, $f(x)$ is strictly positive for $x \neq 0$ and, in consequence, the Taylor series of f about the origin represents f only at the origin! ▶

The example suggests that Taylor approximations may have strong limitations—smoothness, even unlimited smoothness, may not suffice. Faced with this discouraging example, the reader may be tempted to give up on power series. But that would be unfortunate. Polynomial approximations provide powerful analytical tools and we should not give up on them too hastily. An ambitious new question may then be posed: *if f is continuous, not necessarily even differentiable, can it be well approximated by a polynomial?* Karl Weierstrass provided a remarkable answer in a *tour de force* of classical analysis in 1885.⁴

⁴K. Weierstrass, "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen", *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften*

WEIERSTRASS'S POLYNOMIAL APPROXIMATION THEOREM *Let f be a continuous function on the unit interval $[0, 1]$. Then, for any $\epsilon > 0$, there exists a polynomial P (determined by f and ϵ) such that $|P(x) - f(x)| < \epsilon$ for each $x \in [0, 1]$.*

Weierstrass's theorem says in words that f can be *uniformly* well approximated by polynomials. Weierstrass's proof of his theorem flowed from his understanding of Fourier's solution of the heat equation. Many different proofs have since been formulated but among these a little probabilistic gem produced by Sergei N. Bernstein in 1912 continues to captivate and provides a forum for a remarkable application of Chebyshev's inequality.⁵

Suppose f is continuous on the closed unit interval $[0, 1]$. The reader should bear in mind the elementary facts that any continuous function on a closed and bounded interval is uniformly continuous and bounded on that interval (Theorem XXI.2.2 in the Appendix). In particular, there exists M such that $|f(x)| \leq M$ for all $0 \leq x \leq 1$ and, for every $\epsilon > 0$, there exists $\delta = \delta(\epsilon) > 0$, independent of x , such that $|f(y) - f(x)| < \epsilon$ whenever $|y - x| < \delta$.

Now let x be any fixed point in the unit interval $[0, 1]$. Let X_1, X_2, \dots be a sequence of Bernoulli trials with success probability x . Then, for each $n \geq 1$, the sum $S_n = X_1 + \dots + X_n$ has the binomial distribution $P\{S_n = k\} = b_n(k; x) = \binom{n}{k}x^k(1-x)^{n-k}$. By additivity of expectation, $\frac{1}{n}S_n$ has mean x , and its variance likewise is $x(1-x)/n$. As the variance decays monotonically to zero as n increases, the normalised sum $\frac{1}{n}S_n$ is concentrated around the fixed value x . In view of the uniform continuity of f , this suggests that the random variable $f(\frac{1}{n}S_n)$ is likely to take values near $f(x)$. Eliminating uncertainties by taking expectation, we are hence led to consider the expected value

$$f_n(x) := E[f(\frac{1}{n}S_n)] = \sum_j f(\frac{j}{n})b_n(j; x) \quad (5.1)$$

as a possible surrogate for $f(x)$.

As $\sum_j b_n(j; x) = 1$ for any choice of n and x , $f(x) = \sum_j f(x)b_n(j; x)$ and we have

$$|f_n(x) - f(x)| = \left| \sum_j [f(\frac{j}{n}) - f(x)]b_n(j; x) \right| \leq \sum_j |f(\frac{j}{n}) - f(x)|b_n(j; x).$$

On the right, the contribution from summands for which j/n is close to x is small by virtue of the uniform continuity of f while the contribution from summands for which j/n is removed from x is small by virtue of the concentration of $\frac{1}{n}S_n$ around x . This insight suggests that we may be able to partition the sum on the right into the sum of two small terms. To formalise this argument, select

zu Berlin, 1885 (II).

⁵S. N. Bernstein, "Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités", *Comm. Soc. Math. Kharkow*, vol. 13, pp. 1–2, 1912–1913.

any $\epsilon > 0$ and let A_δ denote the set of integers $0 \leq j \leq n$ for which $|\frac{j}{n} - x| \geq \delta$. Then

$$\begin{aligned} |f_n(x) - f(x)| &\leq \sum_{j \in A_\delta^c} |f(\frac{j}{n}) - f(x)| b_n(j; x) + \sum_{j \in A_\delta} |f(\frac{j}{n}) - f(x)| b_n(j; x) \\ &\leq \epsilon \sum_{j \in A_\delta^c} b_n(j; x) + 2M \sum_{j \in A_\delta} b_n(j; x). \end{aligned}$$

For the first sum on the right we have $\sum_{j \in A_\delta^c} b_n(j; x) = P\{|\frac{1}{n}S_n - x| < \delta\} \leq 1$, the crude estimate for the probability sufficing in view of the small multiplying factor ϵ . We need a more delicate touch for the second sum on the right as the multiplying factor $2M$ is not necessarily small. Chebyshev's inequality provides the necessary tool. We recall that $\frac{1}{n}S_n$ has mean x and variance $x(1-x)/n$ to obtain

$$\sum_{j \in A_\delta} b_n(j; x) = P\{|\frac{1}{n}S_n - x| \geq \delta\} \leq \frac{x(1-x)}{n\delta^2}.$$

The bound on the right is still not entirely satisfactory as it depends upon x . We can eliminate the dependence by observing that the quadratic function $q(x) = x(1-x)$ attains its maximum value of $1/4$ at $x = 1/2$. We hence obtain the uniform bound $\sum_{k \in A_\delta} b_n(k; x) \leq 1/(4n\delta^2) \leq \epsilon/(2M)$ if $n \geq M/(2\epsilon\delta^2)$. It follows that $|f_n(x) - f(x)| \leq 2\epsilon$, eventually, for a choice of n determined only by ϵ (and not by x). As $\epsilon > 0$ may be chosen arbitrarily small, the function f_n is a uniformly good pointwise approximation to f for all sufficiently large n .

What is the character of the approximations f_n to f ? It is clear that the sum on the right in (5.1) is only formally infinite as $b_n(j; x)$ is identically zero for $j < 0$ or $j > n$. By setting $h = 1/n$ and rearranging terms in the sum, we obtain

$$\begin{aligned} f_n(x) &= \sum_{j=0}^n f(jh) \binom{n}{j} x^j (1-x)^{n-j} = \sum_{j=0}^n f(jh) \binom{n}{j} x^j \sum_{i=0}^{n-j} (-1)^i \binom{n-j}{i} x^i \\ &\stackrel{(i=k-j)}{=} \sum_{j=0}^n \sum_{k=j}^n (-1)^{k-j} \binom{n}{j} \binom{n-j}{k-j} f(jh) x^k = \sum_{k=0}^n \binom{n}{k} \left\{ \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(jh) \right\} x^k \end{aligned} \tag{5.2}$$

in view of the elementary combinatorial identity $\binom{n}{j} \binom{n-j}{k-j} = \binom{n}{k} \binom{k}{j}$. An examination of the first few summands in the inner sum suggests a compact characterisation of the summands in terms of differences.

For any $h > 0$, we define the *difference operator* $\Delta = \Delta_h$ by

$$\Delta f(x) = f(x+h) - f(x).$$

As a matter of convention, we set $\Delta^0 f(x) = f(x)$, and for $k \geq 1$ define the iterates $\Delta^k f(x) = \Delta(\Delta^{k-1} f(x))$. Then we have

$$\begin{aligned}\Delta^0 f(x) &= f(x), \\ \Delta^1 f(x) &= f(x+h) - f(x), \\ \Delta^2 f(x) &= f(x+2h) - 2f(x+h) + f(x), \\ \Delta^3 f(x) &= f(x+3h) - 3f(x+2h) + 3f(x+h) - f(x), \\ \Delta^4 f(x) &= f(x+4h) - 4f(x+3h) + 6f(x+2h) - 4f(x+h) + f(x),\end{aligned}$$

leading to a combinatorial guess.

LEMMA $\Delta^k f(x) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(x+jh).$

PROOF: It is natural to attempt a proof by induction, the base case already established. It will simplify matters to extend the sum over all integers j as, by convention for binomial coefficients, $\binom{k}{j} = 0$ if $j < 0$ or $j > k$. By induction hypothesis, we now have

$$\begin{aligned}\Delta^{k+1} f(x) &= \Delta(\Delta^k f(x)) = \Delta\left(\sum_j (-1)^{k-j} \binom{k}{j} f(x+jh)\right) \\ &= \sum_j (-1)^{k-j} \binom{k}{j} f(x+h+jh) - \sum_j (-1)^{k-j} \binom{k}{j} f(x+jh) \\ &= \sum_j (-1)^{k+1-j} \left\{ \binom{k}{j-1} + \binom{k}{j} \right\} f(x+jh) = \sum_j (-1)^{k+1-j} \binom{k+1}{j} f(x+jh),\end{aligned}$$

the final step following by Pascal's triangle and completing the induction. ▶

Identifying h with n^{-1} in (5.2), it follows that

$$f_n(x) = \sum_{j=0}^n f\left(\frac{j}{n}\right) \binom{n}{j} x^j (1-x)^{n-j} = \sum_{k=0}^n \binom{n}{k} \Delta^k f(0) x^k. \quad (5.3)$$

We have proved somewhat more than advertised: *the sequence of polynomials $\{f_n, n \geq 0\}$ converges uniformly to the continuous function f on the unit interval $[0, 1]$.* The constructive approximations f_n of the function f are the justly celebrated *Bernstein polynomials*.

6 Some number-theoretic sums

In the next section we shall see a beautiful example of how a probabilistic viewpoint allows one to see number-theoretic arguments in a new light. The context

is a famous theorem of Hardy and Ramanujan on prime factors. We begin with a little number-theoretic preparation in this section.

In this and the following section p and q denote primes (by convention, the first prime is 2), $m \geq 1$ and $n \geq 1$ positive integral variables, $x \geq 1$ a positive real variable, and $\lfloor x \rfloor$ the greatest integer not larger than x (or, equivalently, the integer part of x). Sums with respect to p are implicitly over the entire indicated range of primes; likewise, sums over n are over the entire indicated range of positive integers. Except for stray exceptions introduced for notational reasons, logarithms are always “natural” (to the Napier base e); the logarithm base will be specified explicitly on the few occasions where it is notationally advantageous to switch to a different base. Thus, I will occasionally write $\log_p x = \log x / \log p$ to simplify expressions.

In these contexts asymptotic order notation helps clear the brush. Suppose $f(x)$ and $g(x)$ are two functions. Then: $f = \mathcal{O}(g)$ means that there exists a constant A such that $|f(x)| \leq Ag(x)$ for all x ; $f \asymp g$ means that there exist constants A and B such that $Ag(x) \leq f(x) \leq Bg(x)$ for all x ; $f \sim g$ means that $f(x)/g(x) \rightarrow 1$ as $x \rightarrow \infty$; and $f = o(g)$ means that $f(x)/g(x) \rightarrow 0$ as $x \rightarrow \infty$.

1° THE SUM $R(x) = \sum_{n \leq x} \log n = \log(\lfloor x \rfloor !)$.

As a preamble we begin by considering the logarithm of the factorial. While Stirling’s formula provides ready asymptotic results, an elementary estimation of sums by integrals is both salutary and instructive.

THEOREM 1 (ROBBINS’S BOUNDS FOR THE FACTORIAL) *For every n we have the double inequality*

$$\sqrt{2\pi} n^{n+1/2} e^{-n} \cdot e^{(12n+1)^{-1}} < n! < \sqrt{2\pi} n^{n+1/2} e^{-n} \cdot e^{(12n)^{-1}} \quad (6.1)$$

and, *a fortiori*, $R(x) = x \log x + \mathcal{O}(x)$.

PROOF: As the logarithm increases monotonically we may bound the sum in question by the integrals

$$\int_1^n \log t dt < \sum_{k \leq n} \log k < \int_1^{n+1} \log t dt,$$

or, by an elementary integration by parts,

$$n \log n - n + 1 < R(n) < (n+1) \log(n+1) - n.$$

The double inequality suggests that we split the difference between the upper and lower bounds and compare $R(n)$ to the arithmetic mean of the two. We accordingly consider the difference

$$D(n) = R(n) - \left(n + \frac{1}{2}\right) \log n + n.$$

The Law of Large Numbers

A comparison of successive terms in the difference sequence $\{D(n)\}$ then shows that

$$D(n) - D(n+1) = \left(n + \frac{1}{2}\right) \log\left(\frac{n+1}{n}\right) - 1.$$

By centring the numerator and the denominator of the fraction inside the logarithm to their arithmetic mean, we obtain

$$\frac{n+1}{n} = \frac{n + \frac{1}{2} + \frac{1}{2}}{n + \frac{1}{2} - \frac{1}{2}} = \frac{1 + \frac{1}{2(n+\frac{1}{2})}}{1 - \frac{1}{2(n+\frac{1}{2})}} = \frac{1 + \frac{1}{2n+1}}{1 - \frac{1}{2n+1}}.$$

The Taylor series expansion $\log \frac{1+x}{1-x} = 2(x + \frac{1}{3}x^3 + \frac{1}{5}x^5 + \frac{1}{7}x^7 + \dots)$ for the logarithm hence shows that

$$\begin{aligned} D(n) - D(n+1) &= \frac{2n+1}{2} \log\left(\frac{1 + \frac{1}{2n+1}}{1 - \frac{1}{2n+1}}\right) - 1 \\ &= \frac{1}{3(2n+1)^2} \left[1 + \frac{3}{5} \cdot \frac{1}{(2n+1)^2} + \frac{3}{7} \cdot \frac{1}{(2n+1)^4} + \dots \right]. \end{aligned} \quad (6.2)$$

The terms of the series on the right are all positive and so $D(n) - D(n+1) > 0$ and the sequence $\{D(n)\}$ is decreasing. Now by comparison of the right-hand side of (6.2) with the terms of a geometric series with ratio $(2n+1)^{-2}$ we see that

$$0 < D(n) - D(n+1) < \frac{1}{3(2n+1)^2} \cdot \frac{1}{1 - (2n+1)^{-2}} = \frac{1}{12n} - \frac{1}{12(n+1)}.$$

On the other hand, it is clear that the first term on the right in (6.2) is dominated by the sum and so

$$D(n) - D(n+1) > \frac{1}{3(2n+1)^2} > \frac{1}{12n+1} - \frac{1}{12(n+1)+1},$$

the final step easily verified by clearing fractions. Pooling the bounds, we obtain

$$\frac{1}{12n+1} - \frac{1}{12(n+1)+1} < D(n) - D(n+1) < \frac{1}{12n} - \frac{1}{12(n+1)}.$$

So, on the one hand, the sequence $D'(n) = D(n) - \frac{1}{12n}$ is increasing, while on the other, the sequence $D''(n) = D(n) - \frac{1}{12(n+1)}$ is decreasing. As these two sequences differ termwise by an asymptotically vanishing quantity, $0 < D''(n) - D'(n) = 1/(12n(12n+1)) \rightarrow 0$, it follows that there is a common limit C such that $D'(n) \uparrow C$ and $D''(n) \downarrow C$. We hence see that

$$C + \frac{1}{12n+1} < D(n) < C + \frac{1}{12n}$$

and as $n! = e^{R(n)} = e^{D(n)} n^{n+1/2} e^{-n}$, this shows that

$$e^C n^{n+1/2} e^{-n} \cdot e^{(12n+1)^{-1}} < n! < e^C n^{n+1/2} e^{-n} \cdot e^{(12n)^{-1}}.$$

The ratio of the upper and lower bounds tends to one asymptotically as $n \rightarrow \infty$ and so we obtain $n!/(e^C n^{n+1/2} e^{-n}) \rightarrow 1$. But Stirling's formula then identifies the constant $C = \log \sqrt{2\pi}$. ►

The inequality (6.1) and the beautiful elementary argument are due to H. E. Robbins.⁶ The proof actually gives an alternative derivation of Stirling's formula up to the constant e^C . The constant must then be determined by separate means, say, by the de Moivre–Laplace theorem.

$$2^\circ \text{ THE SUM } \vartheta(x) = \sum_{p \leq x} \log p = \log \prod_{p \leq x} p.$$

This important number-theoretic function is like the previous sum excepting only that we sum over the primes instead.

LEMMA 1 $\vartheta(x) < 2x \log 2$.

PROOF: It is clear that $\vartheta(x) = \vartheta(\lfloor x \rfloor)$ so that it will suffice to consider integer $x = n$. The proof is by induction, the base cases $n = 1$ and $n = 2$ trivial. Suppose the result holds for all $n \leq n_0 - 1$ for some n_0 . If $n_0 = 2m$ is even then

$$\vartheta(n_0) = \vartheta(2m) = \vartheta(2m-1) < 2(2m-1) \log 2 < 2n_0 \log 2.$$

It remains to deal with the case $n_0 = 2m + 1$ odd. Now $\vartheta(2m+1) - \vartheta(m+1) = \log \prod p$ where the product is over all primes in the range $m+1 < p \leq 2m+1$. For p in this range the greatest common divisor of p and $m!$ is one (they are relatively prime) but p divides the integer $(2m+1)^m = (m+1)(m+2) \cdots (2m+1)$. Accordingly, $\prod_{m+1 < p \leq 2m+1} p$ divides the binomial coefficient $\binom{2m+1}{m} = (2m+1)^m/m!$ and *a fortiori* $\prod_{m+1 < p \leq 2m+1} p \leq \binom{2m+1}{m}$. The observation that the terms $\binom{2m+1}{m}$ and $\binom{2m+1}{m+1}$ each appear once in the binomial expansion of $(1+1)^{2m+1}$ shows that $\binom{2m+1}{m} + \binom{2m+1}{m+1} < 2^{2m+1}$ and as $\binom{2m+1}{m+1} = \binom{2m+1}{m}$ it follows that $\binom{2m+1}{m} < 2^{2m}$. Taking logarithms, we obtain $\vartheta(2m+1) - \vartheta(m+1) < 2m \log 2$. It follows that

$$\begin{aligned} \vartheta(n_0) &= \vartheta(2m+1) = \vartheta(m+1) + [\vartheta(2m+1) - \vartheta(m+1)] \\ &< 2(m+1) \log 2 + 2m \log 2 = 2(2m+1) \log 2 = 2n_0 \log 2, \end{aligned}$$

completing the induction. ►

⁶H. E. Robbins, "A remark on Stirling's formula", *American Mathematical Monthly*, vol. 62, pp. 26–29, 1955.

3° THE SUM $\psi(x) = \sum_{p \leq x} \lfloor \log_p x \rfloor \log p$.

The largest value of m for which $p^m \leq x$ is $m = \lfloor \frac{\log x}{\log p} \rfloor = \lfloor \log_p x \rfloor$. For each prime $p \leq x$ the sum for $\psi(x)$ hence includes $\log p$ once for each power of p that does not exceed x . As $p^\ell \leq x$ if, and only if, $p \leq x^{1/\ell}$, we may decompose $\psi(x)$ systematically into the sum

$$\psi(x) = \vartheta(x) + \vartheta(x^{1/2}) + \vartheta(x^{1/3}) + \cdots + \vartheta(x^{1/k})$$

where k is the largest integer value for which $x^{1/k} \geq 2$, which is the same as saying $x \geq 2^k$ or $k = \lfloor \log_2 x \rfloor$. For $2 \leq \ell \leq k$ we have

$$\vartheta(x^{1/\ell}) = \sum_{p \leq x^{1/\ell}} \log p \leq \frac{1}{\ell} x^{1/\ell} \log x \leq x^{1/2} \log x.$$

Pooling the subdominant terms it now follows that

$$\vartheta(x^{1/2}) + \vartheta(x^{1/3}) + \cdots + \vartheta(x^{1/k}) \leq kx^{1/2} \log x = \mathcal{O}(x^{1/2} \log(x)^2).$$

LEMMA 2 $\psi(x) = \vartheta(x) + \mathcal{O}(x^{1/2} \log(x)^2)$.

We now only need a lower bound on $\psi(x)$ to nail down the rates of growth of both $\psi(x)$ and $\theta(x)$. And here it is.

LEMMA 3 $\psi(x) \geq \frac{1}{4}x \log 2$.

PROOF: Any integer $n > 1$ may be expressed in its prime factorisation in the form $n = \prod_p p^{\nu(p)}$ where, for each p , $\nu(p) = \nu_n(p)$ is integral and ≥ 0 . It follows that we may write $n! = \prod_p p^{j_n(p)}$ where, for each p , the integer $j_n(p)$ represents the number of multiples of a power of p that are $\leq n$. In other words, with $m = \lfloor \log_p n \rfloor$,

$$j_n(p) = \left\lfloor \frac{n}{p} \right\rfloor + \left\lfloor \frac{n}{p^2} \right\rfloor + \cdots + \left\lfloor \frac{n}{p^m} \right\rfloor. \quad (6.3)$$

By taking logarithms of the binomial coefficient $\binom{2n}{n} = (2n)!/(n!)^2$, we hence obtain

$$\log \binom{2n}{n} = \sum_p (j_{2n}(p) - 2j_n(p)) \log p = \sum_p \sum_\ell \left(\left\lfloor \frac{2n}{p^\ell} \right\rfloor - 2 \left\lfloor \frac{n}{p^\ell} \right\rfloor \right) \log p.$$

The difference within the round parentheses on the right is equal to 0 if $\lfloor 2n/p^\ell \rfloor$ is even and is equal to 1 if $\lfloor 2n/p^\ell \rfloor$ is odd. As the inner summands vanish for

$\ell > \lfloor \log_p 2n \rfloor = \lfloor \log 2n / \log p \rfloor$, it follows that the contribution from the inner sum on the right is no more than $\lfloor \log_p 2n \rfloor$ and hence

$$\log \binom{2n}{n} \leq \sum_p \lfloor \log_p 2n \rfloor \log p = \psi(2n).$$

On the other hand, we have

$$\binom{2n}{n} = \frac{(2n)n!}{n!} = \frac{(n+1)}{1} \cdot \frac{(n+2)}{2} \cdots \frac{(2n)}{n} \geq 2^n.$$

Taking logarithms again, we obtain $\psi(2n) \geq \log \binom{2n}{n} \geq n \log 2$. Setting $n = \lfloor x/2 \rfloor$, we obtain $\psi(x) \geq \psi(2\lfloor \frac{x}{2} \rfloor) \geq \lfloor \frac{x}{2} \rfloor \log 2 \geq \frac{1}{4}x \log 2$. ►

By pooling Lemmas 1, 2, and 3 we obtain an exquisitely simple result.

THEOREM 2 $\vartheta(x) \asymp x$ and $\psi(x) \asymp x$.

The great utility of these functions is in the proof of the prime number theorem.

4° THE FUNCTION $\pi(x)$.

The growth rate of the prime numbers is captured by the function $\pi(x)$ which stands for the number of primes no larger than x . Now, on the one hand, we have $\vartheta(x) \leq \pi(x) \log x$ by the elementary replacement of the upper bound $\log x$ for each of the terms $\log p$ in the sum for $\vartheta(x)$. On the other hand, by truncating the terms of the sum for $\vartheta(x)$, we have $\vartheta(x) \geq \sum_{x^{1-\delta} < p \leq x} \log p$ where, by replacing each summand in the sum on the right by the lower bound $\log x^{1-\delta}$, the right-hand side is bounded below by

$$(1 - \delta) \log(x)(\pi(x) - \pi(x^{1-\delta})) \geq (1 - \delta) \log(x)(\pi(x) - x^{1-\delta}).$$

Pooling the bounds with Theorem 2, we obtain the beautiful result proved by Chebyshev in 1852.⁷

THEOREM 3 $\pi(x) \asymp x/\log x$.

Somewhat more can be said. Building on the remarkable ideas presented in Bernhard Riemann's 1859 memoir, Hadamard and de la Vallée Poussin independently proved the prime number theorem: $\pi(x) \sim x/\log x$. The argument will take us too far afield and we will be satisfied with Chebyshev's bounds.

5° THE SUM $Q(x) = \sum_{p \leq x} p^{-1} \log p$.

The sum $Q(x)$ is not important in itself but paves the way for the next sum which is.

⁷P. L. Chebyshev, "Mémoire sur les nombres premiers", *Journal de Mathématiques pures et appliquées*, vol. 17, pp. 366–390, 1852.

THEOREM 4 $Q(x) = \log x + \mathcal{O}(1)$.

PROOF: Writing $m = \lfloor \log_p x \rfloor$ in the representation for the factorial in the proof of Lemma 3, we have

$$\log \lfloor x \rfloor! = \sum_{p \leq x} j_n(p) \log p = \sum_{p \leq x} \left(\left\lfloor \frac{x}{p} \right\rfloor + \left\lfloor \frac{x}{p^2} \right\rfloor + \cdots + \left\lfloor \frac{x}{p^m} \right\rfloor \right) \log p. \quad (6.4)$$

Replacing each of the terms $\lfloor xp^{-\ell} \rfloor$ in the round brackets on the right by $xp^{-\ell}$ occasions a correction of no more than one for each term. Consequently, the right-hand side differs from $\sum_{p \leq x} \sum_{\ell \leq m} xp^{-\ell} \log p$ in no more than

$$\sum_{p \leq x} m \log p = \sum_{p \leq x} \lfloor \log_p x \rfloor \log p = \psi(x).$$

By isolating the first term in the sum on the right in (6.4), by Theorem 2 it follows that

$$\log \lfloor x \rfloor! = x \sum_{p \leq x} \frac{\log p}{p} + x \sum_{p \leq x} \left(\frac{1}{p^2} + \cdots + \frac{1}{p^m} \right) \log p + \mathcal{O}(x).$$

Factoring out the term x and identifying the sum $Q(x)$ on the right, we obtain

$$|Q(x) - \frac{1}{x} R(x)| = \sum_{p \leq x} \left(\frac{1}{p^2} + \cdots + \frac{1}{p^m} \right) \log p + \mathcal{O}(1)$$

where $\mathcal{O}(1)$ represents a function of x that is absolutely bounded by some constant. The finite geometric sum in the round brackets on the right is bounded above by $1/p^2(1 - p^{-1}) = 1/p(p - 1)$. It follows that

$$\sum_{p \leq x} \left(\frac{1}{p^2} + \cdots + \frac{1}{p^m} \right) \log p \leq \sum_{p \leq x} \frac{\log p}{p(p - 1)} \leq \sum_n \frac{\log n}{n(n - 1)},$$

the series on the right convergent (by, say, the integral test), and hence $\mathcal{O}(1)$, because the summand decays faster than $n^{-(1+\delta)}$ for any $0 < \delta < 1$. We've shown that $Q(x) = \frac{1}{x} R(x) + \mathcal{O}(1)$ which in view of Theorem 1 is the conclusion to be reached. ▶

6° THE SUM $P(x) = \sum_{p \leq x} p^{-1}$.

It will be convenient to introduce the sequence c_n where $c_p = p^{-1} \log p$ when $n = p$ is prime and $c_n = 0$ if n is not prime. By defining $f(x) = 1/\log x$, we may write the sum $P(x)$ in the form

$$P(x) = \sum_{p \leq x} \frac{\log p}{p} \cdot \frac{1}{\log p} = \sum_{n \leq x} c_n f(n).$$

Now $Q(x) = \sum_{n \leq x} c_n$, or, equivalently, $c_n = Q(n) - Q(n-1)$. Temporarily set $M = \lfloor x \rfloor$ to reduce the notational burden of parentheses. By rearranging terms in the sum we may now write

$$\begin{aligned} P(x) &= Q(2)f(2) + [Q(3) - Q(2)]f(3) + \cdots + [Q(M) - Q(M-1)]f(M) \\ &= Q(2)[f(2) - f(3)] + \cdots + Q(M-1)[f(M-1) - f(M)] + Q(M)f(M). \end{aligned}$$

The function $Q(x)$ has jumps only at integer points (more precisely, only at the primes) and is level in between points of jump. Formally, $Q(x) = Q(n)$ if $n \leq x < n+1$ and *a fortiori* $Q(x) = Q(M)$. This suggests an opportunity to estimate the sum on the right by an integral. By the fundamental theorem of calculus, we have

$$\begin{aligned} P(x) &= - \sum_{n \leq M-1} \int_n^{n+1} Q(t)f'(t) dt + Q(M)f(M) \\ &= - \int_2^M Q(t)f'(t) dt + Q(M)f(M) = - \int_2^x Q(t)f'(t) dt + Q(x)f(x), \quad (6.5) \end{aligned}$$

the final step obtained by subtracting and adding the term $Q(M)[f(x) - f(M)] = Q(x)f(x) - Q(M)f(M)$ in the penultimate term to simplify the integral expression by expanding its range slightly. By Theorem 4, $Q(x)f(x) = 1 + \mathcal{O}(1/\log x) = \mathcal{O}(1)$ while the integral on the right differs from $-\int_2^x \frac{dt}{t \log t}$ in no more than

$$A \int_2^x \frac{dt}{t \log(t)^2} \leq A \int_{\log 2}^{\infty} \frac{du}{u^2} = A \log 2$$

for some constant A . (The change of variable $\log(t) = u$ in the penultimate step is both natural and effective, while expanding the range of the integral can only increase its value as the integrand is positive.) All that's left to evaluate is the elementary integral

$$\int_2^x \frac{dt}{t \log t} = \int_{\log 2}^{\log x} \frac{du}{u} = \log \log x - \log \log 2.$$

Pulling the pieces together in (6.5) we obtain the estimate we seek.

THEOREM 5 $P(x) = \log \log x + \mathcal{O}(1)$.

If the reader looks at the proofs critically she will realise that the estimates may be improved in the characterisation of the behaviour of the order terms but we will not bother with such refinements—the dominant behaviour of these sums suffices for our purposes. The interested reader will find a more searching examination of these topics in Hardy and Wright's classical text.⁸

⁸G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Chapter XXII. Oxford: Oxford University Press, 1979.

7 The dance of the primes

Let $f_n(p)$ denote the fraction of numbers $k \leq n$ that are divisible by a given prime p , in notation, $f_n(p) = \frac{1}{n} \text{card}\{k \leq n : p \text{ divides } k\} = \frac{1}{n} \lfloor n/p \rfloor$. We hence have $\frac{1}{p} - \frac{1}{n} < f_n(p) \leq \frac{1}{p}$ and it follows that $f_n(p) \rightarrow 1/p$ as $n \rightarrow \infty$ uniformly in p . It is natural hence to call $f(p) = 1/p$ the *density* of the prime p . This appears trite, excepting only the impressive-sounding terminology, because all that we are saying is that every p th number is a multiple of p . The observation begins to gather force only when we consider composite numbers. Suppose p and q are primes with densities $f(p) = 1/p$ and $f(q) = 1/q$, respectively. What can we say about the density of the composite number $m = pq$?

In the natural extension of the notation, write $f_n(pq)$ for the fraction of numbers up till n that are divisible by the composite number pq . It follows as before that $\frac{1}{pq} - \frac{1}{n} < f_n(pq) \leq \frac{1}{pq}$ whence, allowing n to tend to infinity, we see that pq has density $f(pq) = 1/(pq)$. And for the *pièce de resistance*, we observe that $f(pq) = f(p)f(q)$. A philistine might say that this makes a great to-do about the elementary observation $1/(pq) = (1/p)(1/q)$ but that would be to miss the point. The probabilist would recognise that *primes are independent* (at least in the sense of Section V.3 that their density satisfies a product law): *primes do indeed appear to play a game of chance!* A pathway is now opened to a *probabilistic* investigation of number-theoretic concepts by exploiting the independence of the primes. A really elegant illustration of the principle was provided by Paul Turán. Here is the setting.

The fundamental theorem of arithmetic tells us that any positive number $n > 1$ may be expressed as a product of prime factors $n = p_1^{\gamma_1} p_2^{\gamma_2} \cdots p_{\omega}^{\gamma_{\omega}}$ where $\omega = \omega(n)$ is the number of *distinct* prime factors of n (not counting multiplicity). The function $\omega(n)$ varies quite irregularly with n —it jumps sharply whenever n hits a highly composite number but drops precipitously to 1 whenever n hits a prime or a power of a prime. Generally speaking, however, we might anticipate that $\omega(n)$ has a tendency to increase with n . (It's not unlike the stock market—periods of gains punctuated by crashes.) The extreme irregularity of the function may suggest however that it would be difficult to capture any “rate of growth”. But, remarkably, G. H. Hardy and S. Ramanujan showed in 1917 that, in a very precise sense, $\omega(n)$ has a very delicate rate of growth.⁹ A definition helps clarify the nature of their result.

We shall say that a function $r(n)$ of the positive integers has *normal order* $s(n)$ if, for every $\epsilon > 0$, we have $(1 - \epsilon)s(n) < r(n) < (1 + \epsilon)s(n)$ for almost all n . Here “almost all” is intended to mean that the size of the exceptional set of integers violating the pair of inequalities grows slowly; more precisely, $\frac{1}{n} \text{card}\{k \leq n : |r(k) - s(k)| \geq \epsilon s(k)\} \rightarrow 0$ as $n \rightarrow \infty$ for every $\epsilon > 0$. With this

⁹G. H. Hardy and S. Ramanujan, “The normal number of prime factors of a number”, *Quarterly Journal of Mathematics*, vol. 48, pp. 76–92, 1917.

for preparation, here is the Hardy–Ramanujan theorem without further ado.

THEOREM *The number of prime factors $\omega(n)$ has normal order $\log \log n$.*

One could hardly wish for a cleaner or simpler result though the original proof is anything but simple. Seventeen years after Hardy and Ramanujan’s paper appeared, Paul Turán provided a beautiful probabilistic proof which can fit within one page! Here it is.¹⁰

PROOF: Fix any $n > 1$ and select an integer N randomly (that is to say, uniformly) from the collection $\{1, \dots, n\}$. For each prime $p \leq n$ let X_p be the indicator for the event that p divides N . Then X_p represents a Bernoulli trial with success probability $\frac{1}{n} \text{ card}\{k \leq n : p \text{ divides } k\} = f_n(p)$, whence $E(X_p) = f_n(p)$ and $\text{Var}(X_p) = f_n(p)(1 - f_n(p))$. In view of our discovery that the density of primes satisfies a product law, we anticipate that if p and q are distinct primes $\leq n$ then X_p and X_q are “asymptotically independent”. More precisely, as $X_p X_q = 1$ if, and only if, pq divides N , we have

$$\begin{aligned} \text{Cov}(X_p, X_q) &= E(X_p X_q) - E(X_p)E(X_q) = f_n(pq) - f_n(p)f_n(q) \\ &< \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) < \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right), \end{aligned}$$

and, as the covariance vanishes when $n \rightarrow \infty$, the random variables X_p and X_q are at least asymptotically uncorrelated.

The number of distinct prime divisors of N is the random variable $\omega(N) = \sum_{p \leq n} X_p$. The stage is set for Chebyshev’s inequality. By additivity of expectation,

$$E(\omega(N)) = \sum_{p \leq n} E(X_p) = \sum_{p \leq n} \frac{1}{p} + O\left(\frac{\pi(n)}{n}\right) = \log \log n + O(1) + O\left(\frac{1}{\log n}\right),$$

by Theorems 3 and 5 of the previous section. As $\pi(n)/n = O(1/\log n)$ vanishes slowly but inexorably, and so is certainly bounded by a constant, we obtain $|E(\omega(N)) - \log \log n| = O(1)$. To estimate the variance, another application of additivity shows that

$$\begin{aligned} \text{Var}(\omega(N)) &= \sum_{p \leq n} \text{Var}(X_p) + \sum_{\substack{p,q \leq n \\ p \neq q}} \text{Cov}(X_p, X_q) < \sum_{p \leq n} \frac{1}{p} + \sum_{\substack{p,q \leq n \\ p \neq q}} \frac{1}{n} \left(\frac{1}{p} + \frac{1}{q}\right) \\ &= \sum_{p \leq n} \frac{1}{p} + \frac{2(\pi(n) - 1)}{n} \sum_{p \leq n} \frac{1}{p} = \log \log n + O(1) + O\left(\frac{\log \log n}{\log n}\right) \end{aligned}$$

¹⁰P. Turán, “On a theorem of Hardy and Ramanujan”, *Journal of the London Mathematical Society*, vol. 9, pp. 274–276, 1934.

where the second-order term on the right decreases quietly to zero and is hence certainly bounded by a constant. The standard deviation of $\omega(n)$ hence increases no faster than the square-root of the mean. We've seen that replacing $E(\omega(N))$ by $\log \log n$ occasions a correction term of at most a constant, and so, for any $\epsilon > 0$, Chebyshev's inequality shows that

$$P\left\{|\omega(N) - \log \log n| > \sqrt{\frac{1}{\epsilon} \log \log n}\right\} \leq \frac{\log \log n + O(1)}{\left(\sqrt{\frac{1}{\epsilon} \log \log n} + O(1)\right)^2} = \epsilon + o(1)$$

where $o(1)$ denotes a function of n which vanishes asymptotically. ▶

Of course, one can construct a purely analytical proof along these lines devoid of probabilistic language. But the analytical manoeuvres would surely appear mysterious to the reader. Sans the probabilistic intuition would such a proof have been discovered? Perhaps. But I doubt it.

Our proof shows somewhat more than advertised: *most numbers n have $\log \log n + O(\sqrt{\log \log n})$ prime factors.* The theorem shows that the number of distinct prime factors is not only not much more than $\log \log n$ but also not much less than $\log \log n$. Thus, most numbers of the order of 10^{80} would have about 5 or 6 prime factors. The reader may be surprised at how few prime factors most large numbers have but what is really surprising is that they have so *many* prime factors; a little introspection will serve to convince the reader that, by the pigeon-hole principle,¹¹ most numbers cannot have too many prime factors—the lower bound is the sticking point in the proof and what surprises.

A number is called “round” if it is the product of a lot of small primes. Experience tells us that round numbers are relatively rare. Our theorem tells us why. If we enumerate the primes in order and let the round number $n_k = p_1 p_2 \cdots p_k$ be the product of the first k primes then $\omega(n_k) = k$. On the other hand, Theorem 3 of the previous section tells us that p_k is of the order of $k \log k$ whence $\log n_k$ is of order $\vartheta(k \log k) = \sum_{p \leq k \log k} \log p = O(k \log k)$. Thus, a typical number of the same order of magnitude as n_k would have approximately $\log \log n_k = O(\log k)$ distinct prime factors which is much much less than k . *Round numbers are atypical.*

The prime factorisation of integers is not only of a fundamental theoretical interest but has become, in the age of the internet, an essential tool of commerce. The protection of private information such as credit card numbers or personal identification from eavesdroppers is now routinely accomplished by embedding internet transactions within the *Hyper Text Transfer Protocol Secure* (or [https](https://), in short) via *public key cryptography*. In such transactions information is cryptographically scrambled using a very large composite number

¹¹If there are m pigeons and n pigeon-holes and $n < m$ then each pigeon cannot get its own pigeon-hole; or, if A and B are finite sets with $\text{card } B < \text{card } A$ then there exists no injection $f: A \rightarrow B$ from A into B .

known to the general public (hence, public key) but whose two large prime factors are known only to the trusted party, say, an internet retailer, who is securing the transactions. Descrambling the information to recover the original message (within a reasonable amount of time) requires knowledge of the prime factors of the public key. The protection of sensitive information within the `https` protocol relies upon the fundamental computational intractability of efficiently factoring very large composite numbers into large constituent prime factors.

8 Fair games, the St. Petersburg paradox

Suppose a sequence of independent, positive random variables X_1, X_2, \dots represents the payouts in repeated trials of a game of chance. If a gambler plays the game n times his accumulated winnings are $S_n = X_1 + \dots + X_n$. If the game has an expected payout of μ per trial it is natural to imagine that the game would be fair if the gambling house collects an entrance fee of μ from the gambler each time he plays the game. Then, after playing the game n times, he will have paid μn in *accumulated entrance fees* and will have won S_n . A theoretical basis for saying that this is a “fair” game is provided by the weak law of large numbers as $\frac{1}{n} S_n \rightarrow^P \mu$ or, equivalently, $P\{|S_n - \mu n| > \epsilon n\} \rightarrow 0$ for every $\epsilon > 0$ or, yet again, $P\left\{\left|\frac{S_n}{\mu n} - 1\right| > \epsilon\right\} \rightarrow 0$ for every $\epsilon > 0$. The final formulation can be used as a basis for defining a fair game even if the expected payout per trial is infinite.

DEFINITION We say that a game with accumulated entrance fees $\{\alpha_n, n \geq 1\}$ is *fair in the classical sense* if $P\left\{\left|\frac{S_n}{\alpha_n} - 1\right| > \epsilon\right\} \rightarrow 0$ for every $\epsilon > 0$. We call any such sequence α_n an (*accumulated*) *fair entrance fee*.

The reader should be cautioned that fairness in the classical sense just means that a limit law of a particular character is in force. It does *not* mean that the gambler will, on average, break even in a long sequence of trials. Indeed it is possible to construct classically fair games in which the gambler is *guaranteed* to lose money. The reader will find an example in the *Problems*. The moral? Look at the fine print. The father’s advice to his son in D. Runyon’s *The Idyll of Miss Sarah Brown* captures the sentiment well.

‘Son,’ the old guy says, ‘no matter how far you travel, or how smart you get, always remember this. Someday, somewhere,’ he says, ‘a guy is going to come to you, and show you a nice brand new set of cards on which the seal is never broken, and this guy is going to offer to bet you that a jack of spades will jump out of the deck and squirt cider in your ear. But son,’ the old guy says, ‘do not bet him, for as sure as you do you are going to get an earful of cider.’

The archetypal game without expectation was proposed by Nicolas Bernoulli in 1713 in a letter to de Montmort. The game was presented by Nicolas’s cousin Daniel Bernoulli in the *Commentaries of the Imperial Academy of Sciences of Saint Petersburg* in 1738, hence the moniker *St. Petersburg paradox*.

Consider a game in which a coin is tossed repeatedly until the first head is observed. The payout X of the game is determined by how many tosses it took to observe the head: if the first head occurred on the k th toss then the payout is 2^k . Of course, the

probability of this is 2^{-k} . The expected payout per trial then is $E(X) = \sum_{k \geq 1} 2^k \cdot 2^{-k} = \infty$. Thus, while half the time the payout is \$2 and three-quarters of the time the payout is no more than \$4, the expected payout is infinite and the gambler should be willing to pay any fixed entrance fee, however large. In practice, many commentators would agree with Nicolas Bernoulli's own assessment that "There ought not to exist any even halfway sensible person who would not sell the right of playing the game for 40 ducats." While explanations of this apparent paradox have been proposed based on risk-aversion and marginal utility, as shown by W. Feller,¹² a perfectly reasonable solution can be crafted in terms of accumulated entrance fees.

Suppose α_n is an accumulated entrance fee to be determined for the St. Petersburg game. The problem sets up nicely for the method of truncation. As in the proof of Khinchin's law of large numbers, for each j , set $U_j = X_j 1_{(-\infty, \alpha_n]}(X_j)$ and $V_j = X_j 1_{(\alpha_n, \infty)}(X_j)$ where $\{\alpha_n, n \geq 1\}$ is a divergent positive sequence to be determined. As $X_j = U_j + V_j$ for each j , the triangle inequality yields

$$|X_1 + \cdots + X_n - \alpha_n| \leq |U_1 + \cdots + U_n - \alpha_n| + |V_1 + \cdots + V_n|.$$

By Boole's inequality it hence follows that

$$\begin{aligned} P\{|S_n - \alpha_n| > \epsilon \alpha_n\} &\leq P\{|U_1 + \cdots + U_n - \alpha_n| > \frac{1}{2} \epsilon \alpha_n\} \\ &\quad + P\{|V_1 + \cdots + V_n| > \frac{1}{2} \epsilon \alpha_n\} \end{aligned} \quad (8.1)$$

as the occurrence of the event on the left implies the occurrence of at least one of the events on the right. A crude bound suffices to estimate the second term on the right. Now, if the event $|V_1 + \cdots + V_n| > \epsilon \alpha_n / 2$ occurs then certainly at least one of the V_j must be non-zero. By another application of Boole's inequality we hence have

$$P\{|V_1 + \cdots + V_n| > \frac{1}{2} \epsilon \alpha_n\} \leq \sum_{j=1}^n P\{|X_j| > \alpha_n\} = n \sum_{k > \log_2 \alpha_n} 2^{-k} \leq \frac{2n}{\alpha_n}, \quad (8.2)$$

and for the bound to be useful we must have $n \ll \alpha_n$.

To estimate the first term on the right of (8.1) we observe that the truncated variables U_j take values in a bounded interval and hence have finite expectation and variance. We have

$$\begin{aligned} E(U_j) &= \sum_{k \leq \log_2 \alpha_n} 2^k 2^{-k} = \lfloor \log_2 \alpha_n \rfloor, \\ \text{Var}(U_j) &\leq E(U_j^2) = \sum_{k \leq \log_2 \alpha_n} 2^{2k} 2^{-k} = 2^{\lfloor \log_2 \alpha_n \rfloor + 1} - 1 < 2\alpha_n. \end{aligned}$$

As the sum $\frac{1}{n}(U_1 + \cdots + U_n)$ will concentrate at $E(U_j)$ we should choose the accumulated entrance fee in the first term on the right in (8.1) to satisfy $\alpha_n \sim n \log_2 \alpha_n$. (For an explanation of the asymptotic order notation see the second paragraph of Section 6.) By Chebyshev's inequality we then have

$$P\{|U_1 + \cdots + U_n - \alpha_n| > \frac{1}{2} \epsilon \alpha_n\} \leq \frac{8n\alpha_n}{\epsilon^2 \alpha_n^2} \quad (8.3)$$

¹²W. Feller, *An Introduction to Probability Theory and Its Applications, Volume I*, pp. 251–253. New York: John Wiley & Sons, 1968.

and we will require $n \ll \alpha_n^2/\alpha_n$ or $n \gg \alpha_n/\log_2(\alpha_n)^2$.

To figure out the proper truncation, it is easiest to equate the rates of decay of the bounds of (8.2) and (8.3). We find that we should select α_n so that $\alpha_n^2/\log_2(\alpha_n)^2 = n^2$ or $\alpha_n \sim n \log_2 n$. With such a selection, both bounds decay as $1/\log_2 n$ and we have determined an accumulated fair entrance fee.

THEOREM *The sequence $\alpha_n = n \log_2 n$ is a fair entrance fee for the St. Petersburg game.*

The marginal, per game cost increases slowly each time the game is played.

The nature of the game changes if the gambler is allowed to purchase shares in the game and elects to reinvest some or all of his winnings in each trial. This setting can be analysed via the weak law as well and is outlined in the *Problems*.

9 Kolmogorov's law of large numbers

The weak law of large numbers is not entirely satisfactory. While it is true that it gives us some confidence in a “law of averages”, the guarantees it provides are of the most meagre. To understand why this is so we will need to revisit the ideas of convergence in probability and convergence a.e.

Suppose X_1, X_2, \dots are independent random variables drawn from a common distribution F with finite mean μ . As usual, let $S_n = X_1 + \dots + X_n$ and consider the weighted sequence of random sums $\frac{1}{n}S_n$. A given sample point ω engenders a particular realisation of these values, the fixed real sequence $\{\frac{1}{n}S_n(\omega), n \geq 1\}$ called a *sample path*. If we imagine that the index n represents the passage of time then a given sample path traces through an evolution of values in time and the weak law says that at any given, sufficiently large, epoch in time n , the values $\frac{1}{n}S_n$ of most sample paths are clustered near the value μ . Unfortunately, however, for any given sample point ω , the fact that at a given epoch in time $\frac{1}{n}S_n(\omega)$ is close to μ gives no guarantees that as one progresses along the sample path *all* subsequent values $\frac{1}{k}S_k(\omega)$ ($k \geq n$) will continue to remain close to μ . That is to say, a sample path which is “good” at epoch n (in the sense that its value at n is close to μ) may become aberrant at a later epoch n' (in the sense that its value at n' has drifted away from μ). Indeed, there is no guarantee that any given sample path will not become bad infinitely often at epochs in time. All that the weak law allows us to conclude is that there is a majority of sample paths that are good at n and a majority of sample paths that are good at n' though the composition of what constitutes the majority can conceivably change from epoch to epoch.

This is unfortunate. One only has to think on how disquieting the situation would be if empirical estimates of the sizes of various subpopulations in a given population actually *worsen* if we increase the size of the random sample in a poll from n to n' . If our intuitive empirically driven ideas of probability are to gain any traction at all one surely expects (or perhaps it is fairer to say that

one hopes!) that increasing the sample size can only improve the estimate. Or, translated into notation, we expect that, for most sample points ω , the values $\frac{1}{n}S_n(\omega)$ converge as a real sequence to μ . And indeed Kolmogorov was able to strengthen Khinchin's law of large numbers to show that this is the case.¹³

THE STRONG LAW OF LARGE NUMBERS *Suppose X_1, X_2, \dots is a sequence of independent random variables drawn from a common distribution with finite mean μ and let $S_n = X_1 + \dots + X_n$ for each n . Then $\frac{1}{n}S_n \rightarrow^{a.e.} \mu$ or, spelled out, $\mathbf{P}\left\{\left|\frac{1}{n}S_n - \mu\right| \geq \epsilon \text{ i.o.}\right\} = 0$ for every $\epsilon > 0$.*

The strong law is a crown jewel of probability. It closes the loop with our intuitive frequency-based ideas of probability and validates at least the principle behind the polling mechanisms that are so ubiquitous today: the relative frequency of occurrence of any event in a sequence of independent trials does indeed converge (a.e.) to the event probability.

The basic idea behind the proof is to have Chebyshev's inequality punch above its weight by coupling it with the Borel–Cantelli lemma. For this to work the tails of the distribution have to die quickly enough that the Borel–Cantelli lemma can come to the fore. This was precisely the situation in our earlier exposure to the strong law in the particular context of Borel's law of normal numbers. If the reader revisits Section V.7 and reconstructs the proof with the benefit of experience she will realise that the proof exploited the existence of a fourth moment to show that the distribution tails died sufficiently quickly. From this understanding it becomes clear that the proof given there in essence can be extended to any sequence of random variables with a fourth moment.

In order to prove that nothing beyond the mean is needed requires correspondingly more dexterity. In the proof of his theorem Kolmogorov introduced two productive new devices, *maximal inequalities* and *moving truncations*, whose utility far transcends the immediate purposes of the theorem. Both ideas repay careful consideration.

It will be useful to expand consideration to a slightly more general setting where the summands may have different distributions. Accordingly, suppose X_1, X_2, \dots is a sequence of independent, square-integrable random variables and $S_n = X_1 + \dots + X_n$ the corresponding sequence of partial sums. Write $\mu_j = \mathbf{E}(X_j)$ and $\sigma_j^2 = \text{Var}(X_j)$, and set $m_n = \mu_1 + \dots + \mu_n$ and $s_n^2 = \sigma_1^2 + \dots + \sigma_n^2$.

KOLMOGOROV'S MAXIMAL INEQUALITY $\mathbf{P}\left\{\max_{1 \leq j \leq n} |S_j - m_j| \geq t\right\} \leq s_n^2/t^2$.

PROOF: Writing $A = \bigcup_{j=1}^n \{|S_j - m_j| \geq t\}$ for the event on the left, it is clear that if A occurs then there must be a *smallest* index v for which $|S_v - m_v| \geq t$ and

¹³Kolmogorov established sufficient conditions for variables with variance in 1930: A. N. Kolmogorov, "Sur la loi forte des grands nombres", *C.R. Acad. Sci. Paris Sér. I Math.*, vol. 191, pp. 910–912, 1930. Necessary and sufficient conditions came in 1933: A. N. Kolmogorov, *Foundations of the Theory of Probability, op. cit.*

we are hence led to consider the events

$$A_v = \{ |S_1 - m_1| < t, \dots, |S_{v-1} - m_{v-1}| < t, |S_v - m_v| \geq t \}$$

and their indicators 1_{A_v} . As the events A_1, \dots, A_n are mutually exclusive and their union is A it is clear that $P(A) = \sum_{v=1}^n P(A_v) = \sum_{v=1}^n E(1_{A_v})$. Matters have set up beautifully for Chebyshev's device. On the set of sample points determined by A_v we have $|S_v - m_v| \geq t$ and so $1_{A_v} \leq 1_{A_v}(S_v - m_v)^2/t^2$. By monotonicity of expectation it follows that $E(1_{A_v}) \leq E(1_{A_v}(S_v - m_v)^2)/t^2$. We may relate the centred partial sum on the right to the centred sum of the entire sequence by the trite but effective manoeuvre $S_n - m_n = (S_v - m_v) + [(S_n - m_n) - (S_v - m_v)] = (S_v - m_v) + S'_{n-v}$ where $S'_{n-v} = \sum_{j=v+1}^n (X_j - \mu_j)$ is the centred partial sum of the *reversed* sequence. By expanding out the square it follows by additivity of expectation that

$$\begin{aligned} E(1_{A_v}(S_n - m_n)^2) &= E(1_{A_v}(S_v - m_v)^2) \\ &\quad + E(1_{A_v} S'^2_{n-v}) + 2E(1_{A_v}(S_v - m_v)S'_{n-v}). \end{aligned}$$

As A_v and $S_v - m_v$ are completely determined by the random variables X_1, \dots, X_v , and are hence independent of X_{v+1}, \dots, X_n , it follows that S'_{n-v} is independent of both 1_{A_v} and $S_v - m_v$. And *a fortiori* the random variables S'_{n-v} and $1_{A_v}(S_v - m_v)$ are uncorrelated. As S'_{n-v} has zero mean the final term on the right of the previous equation hence vanishes and we are left with

$$\begin{aligned} E(1_{A_v}(S_n - m_n)^2) &= E(1_{A_v}(S_v - m_v)^2) + E(1_{A_v} S'^2_{n-v}) \\ &\geq E(1_{A_v}(S_v - m_v)^2) \geq t^2 E(1_{A_v}). \end{aligned}$$

It only remains to work back and appeal once more to additivity to obtain

$$\begin{aligned} P(A) &= \sum_{v=1}^n E(1_{A_v}) \leq \sum_{v=1}^n \frac{1}{t^2} E(1_{A_v}(S_n - m_n)^2) = \frac{1}{t^2} E\left((S_n - m_n)^2 \sum_{v=1}^n 1_{A_v}\right) \\ &= \frac{1}{t^2} E((S_n - m_n)^2 1_A) \leq \frac{1}{t^2} E((S_n - m_n)^2) \end{aligned}$$

which completes the proof. We have actually proved a little more than asserted and the penultimate inequality is occasionally useful in its own right. ►

The maximal inequality has a superficial resemblance to Chebyshev's inequality but is much much stronger. For example, stitching together n applications of Chebyshev's inequality via Boole's inequality would only give

$$P\left(\bigcup_{1 \leq j \leq n} \{|S_j - m_j| \geq t\}\right) \leq \sum_{j=1}^n P\{|S_j - m_j| \geq t\} \leq \sum_{j=1}^n \frac{s_j^2}{t^2}.$$

The power of Kolmogorov's maximal inequality is immediately apparent in the following useful criterion for a.e. convergence in terms of the variances of the summands.

KOLMOGOROV'S CRITERION *If $\sum_{j=1}^{\infty} \sigma_j^2/j^2$ converges then $\frac{1}{n}(S_n - m_n) \rightarrow^{a.e.} 0$.*

PROOF: Fix any $\epsilon > 0$. By Theorem XII.5.2 we should attempt to show that $P(\bigcup_{j \geq n} \{|\frac{1}{j}(S_j - m_j)| \geq \epsilon\}) \rightarrow 0$. It is natural to attempt to apply Boole's inequality with repeated applications of Chebyshev's inequality to estimate the probability but this naïve attempt yields much too weak a bound. The trick is to break up the union into a succession of larger and larger blocks on each of which Kolmogorov's maximal inequality can be applied. For each $k \geq 1$ define the event $B_k = \bigcup_{2^{k-1} < j \leq 2^k} \{|\frac{1}{j}(S_j - m_j)| \geq \epsilon\}$. Then

$$P\left(\bigcup_{j > 2^{r-1}} \{|\frac{1}{j}(S_j - m_j)| \geq \epsilon\}\right) \leq \sum_{k \geq r} P(B_k)$$

and the right-hand side is the tail of the series $\sum_k P(B_k)$. The series tail vanishes as $r \rightarrow \infty$ if, and only if, the parent series converges and we are hence led to consider the convergence of $\sum_k P(B_k)$. By leveraging Kolmogorov's maximal inequality applied to each B_k , we have

$$\begin{aligned} P(B_k) &= P\left(\bigcup_{2^{k-1} < j \leq 2^k} \{|\frac{1}{j}(S_j - m_j)| \geq \epsilon\}\right) \\ &\leq P\left(\bigcup_{2^{k-1} < j \leq 2^k} \{|S_j - m_j| \geq \epsilon 2^{k-1}\}\right) \leq \frac{s_{2^k}^2}{\epsilon^2 2^{2k-2}} = \frac{4}{\epsilon^2} \sum_{j=1}^{2^k} \sigma_j^2 4^{-k}. \end{aligned}$$

Summing over k and performing the natural interchange of the order of summation we then obtain

$$\begin{aligned} \sum_{k=1}^{\infty} P(B_k) &\leq \frac{4}{\epsilon^2} \sum_{k=1}^{\infty} \sum_{j=1}^{2^k} \sigma_j^2 4^{-k} = \frac{4}{\epsilon^2} \sum_{j=1}^{\infty} \sigma_j^2 \sum_{k \geq \log_2 j} 4^{-k} \\ &\leq \frac{4}{\epsilon^2} \sum_{j=1}^{\infty} \frac{\sigma_j^2 4^{-\log_2 j}}{1 - 4^{-1}} = \frac{16}{3\epsilon^2} \sum_{j=1}^{\infty} \frac{\sigma_j^2}{j^2} \end{aligned}$$

and the series on the right converges by assumption. It follows that the tails of the series $\sum_k P(B_k)$ converge nicely to zero. ▶

If the variables have the same variance $\sigma_j^2 = \sigma^2$ then, in view of the convergence of the series $\sum_{j \geq 1} 1/j^2$, the criterion applies trivially and so this already gives us the strong law if the summands have a common variance. But

we have bigger game in our sights. In order to extend the result further to variables which may not have variance Kolmogorov introduced a new device, the idea of *moving truncations*, to reduce the sequence $\{X_j, j \geq 1\}$ to an equivalent sequence $\{U_j, j \geq 1\}$ to which the criterion can be applied.

DEFINITION We say that two sequences of random variables, $\{X_j, j \geq 1\}$ and $\{U_j, j \geq 1\}$, are *equivalent* if $P\{U_j \neq X_j \text{ i.o.}\} = 0$.

Suppose now that X is an integrable random variable with d.f. F and mean $E(X) = \int_{\mathbb{R}} x dF(x) = \mu$. Then $E(|X|) = M$ is finite and, in view of Theorem XIII.5.4, the series $\sum_{j=1}^{\infty} \int_{|x| \geq j} dF(x) = \sum_{j=1}^{\infty} P\{|X| \geq j\} \leq E(|X|) = M$ is convergent. This suggests a particularly effective moving truncation. Suppose that $\{X_j, j \geq 1\}$ is a sequence of independent random variables drawn from the common distribution F . For each j , define the truncated variable $U_j = X_j 1_{[-j, j]}(X_j)$ which equals X_j when $|X_j| \leq j$ and is zero otherwise. Then

$$\begin{aligned} P\{U_j \neq X_j \text{ i.o.}\} &= \lim_n P\left(\bigcup_{j \geq n} \{U_j \neq X_j\}\right) = \lim_n P\left(\bigcup_{j \geq n} \{|X_j| > j\}\right) \\ &\leq \lim_n \sum_{j \geq n} P\{|X_j| > j\} \leq \lim_n \sum_{j \geq n} P\{|X_j| \geq j\} = \lim_n \sum_{j \geq n} \int_{|x| \geq j} dF(x) = 0 \quad (9.1) \end{aligned}$$

as the series $\sum_{j \geq 1} \int_{|x| \geq j} dF(x)$ is convergent so that its tails must vanish. It follows that the sequences $\{X_j, j \geq 1\}$ and $\{U_j, j \geq 1\}$ are equivalent.

Each U_j is bounded and has moments of all orders. In particular, let $\mu_j = E(U_j)$ and $\sigma_j^2 = \text{Var}(U_j)$. Then

$$\sigma_j^2 \leq \int_{|x| \leq j} x^2 dF(x) = \sum_{k=1}^j \int_{k-1 < |x| \leq k} x^2 dF(x) \leq \sum_{k=1}^j k \int_{k-1 < |x| \leq k} |x| dF(x).$$

Writing $a_k = \int_{k-1 < |x| \leq k} |x| dF(x)$ and summing over j , it follows that

$$\sum_{j=1}^{\infty} \frac{\sigma_j^2}{j^2} \leq \sum_{j=1}^{\infty} \frac{1}{j^2} \sum_{k=1}^j k a_k = \sum_{k=1}^{\infty} k a_k \sum_{j=k}^{\infty} \frac{1}{j^2}.$$

The inner sum on the right is easy to estimate by elementary methods. For instance,

$$\sum_{j=k}^{\infty} \frac{1}{j^2} = \frac{1}{k^2} + \sum_{j=k+1}^{\infty} \frac{1}{j^2} \leq \frac{1}{k^2} + \int_k^{\infty} \frac{dx}{x^2} = \frac{1}{k^2} + \frac{1}{k} \leq \frac{2}{k},$$

the upper bound of the sum by the integral justified as the function $1/x^2$ is monotone. It follows that the series

$$\sum_{j=1}^{\infty} \frac{\sigma_j^2}{j^2} \leq 2 \sum_{k=1}^{\infty} k a_k = 2 \sum_{k=1}^{\infty} \int_{k-1 < |x| \leq k} |x| dF(x) = 2 \int_{|x| > 0} |x| dF(x) = 2M$$

converges and Kolmogorov's criterion is satisfied. Consequently,

$$\frac{1}{n}((U_1 + \cdots + U_n) - (\mu_1 + \cdots + \mu_n)) \rightarrow^{\text{a.e.}} 0. \quad (9.2)$$

The arithmetic means $\bar{\mu}_n = \frac{1}{n}(\mu_1 + \cdots + \mu_n)$ are called the *Cesàro means* of the original sequence $\{\mu_j, j \geq 1\}$. Now it is certainly intuitive that averaging can only improve the behaviour of a sequence so that the sequence $\{\bar{\mu}_n, n \geq 1\}$ should be at least as well behaved as the parent sequence $\{\mu_j, j \geq 1\}$. This is the venerable result of Cesàro. I have included the elementary proof though the reader may well find that the form of the argument is becoming routine.

THE LEMMA OF CESÀRO MEANS *If the sequence $\{\mu_n, n \geq 1\}$ converges then so does the corresponding sequence of Cesàro means $\{\bar{\mu}_n, n \geq 1\}$ and to the same limit. There exist convergent sequences of Cesàro means, however, for which the original sequence does not converge.*

PROOF: Suppose $\mu_n \rightarrow a$ for some a . Fix any tiny, strictly positive ϵ . By the definition of convergence there then exists a positive integer n_0 such that $|\mu_n - a| < \epsilon$ for all $n > n_0$. Write $A = \max\{|\mu_1|, \dots, |\mu_{n_0}|, \dots\}$. As $\bar{\mu}_n - a = \frac{1}{n} \sum_{k=1}^n (\mu_k - a)$, for all $n > n_0$, it follows courtesy the triangle inequality that

$$\begin{aligned} |\bar{\mu}_n - a| &= \frac{1}{n} |(\mu_1 - a) + \cdots + (\mu_n - a)| \\ &\leq \frac{1}{n} (|\mu_0 - a| + \cdots + |\mu_{n_0} - a|) + \frac{1}{n} (|\mu_{n_0+1} - a| + \cdots + |\mu_n - a|). \end{aligned}$$

The summands $|\mu_{n_0+1} - a|, \dots, |\mu_n - a|$ of the second term on the right are each less than ϵ by design. We can afford to be more cavalier in bounding the summands in the first term. As $|\mu_k - a| \leq |\mu_k| + |a|$ by another application of the triangle inequality, each of the summands $|\mu_1 - a|, \dots, |\mu_{n_0} - a|$ in the first term is bounded above by $A + |a|$. Accordingly, we obtain $|\bar{\mu}_n - a| \leq \frac{n_0}{n} (A + |a|) + \frac{n-n_0}{n} \epsilon$. The first term on the right may be made as small as desired, say $< \epsilon$, by choosing n sufficiently large, while the second term on the right is certainly no larger than ϵ . It follows that $|\bar{\mu}_n - a| < 2\epsilon$, eventually. But as this holds for every choice of $\epsilon > 0$ we must have $\bar{\mu}_n \rightarrow a$, as advertised. The oscillating sequence $\mu_n = (-1)^n$ for which $\bar{\mu}_n \rightarrow 0$ shows that the converse of the statement need not hold. ▶

It only remains to show that the sequence $\{\mu_j, j \geq 1\}$ converges. It is clear on the one hand that the sequence $\{U_j, j \geq 1\}$ converges pointwise to X and on the other that $|U_j| \leq |X|$ for each j so that each U_j is dominated absolutely by the integrable random variable $|X|$. By the dominated convergence theorem it follows directly that $\mu_j = E(U_j) \rightarrow E(X) = \mu$. By the lemma of Cesàro means it follows that

$$\bar{\mu}_n = \frac{1}{n}(\mu_1 + \cdots + \mu_n) \rightarrow \mu. \quad (9.3)$$

Pooling (9.1,9.2,9.3) completes the proof of the strong law. Spelled out, from (9.1), for all ω , excepting only points in a null set \mathfrak{N}_1 , we have

$$\left| \frac{1}{n}(X_1(\omega) + \cdots + X_n(\omega)) - \frac{1}{n}(U_1(\omega) + \cdots + U_n(\omega)) \right| < \epsilon,$$

eventually, for all sufficiently large n . From (9.2), for all ω , excepting only points in a null set \mathfrak{N}_2 , we have

$$\left| \frac{1}{n} (U_1(\omega) + \cdots + U_n(\omega)) - \frac{1}{n} (\mu_1 + \cdots + \mu_n) \right| < \epsilon,$$

eventually. Finally, from (9.3),

$$\left| \frac{1}{n} (\mu_1 + \cdots + \mu_n) - \mu \right| < \epsilon,$$

eventually. All three inequalities will hold simultaneously for all $\omega \notin \mathfrak{N}_1 \cup \mathfrak{N}_2$ and a sufficiently large choice of n . By two applications of the triangle inequality, it follows that $\left| \frac{1}{n} S_n(\omega) - \mu \right| < 3\epsilon$, eventually, for all ω excepting only on a null set of zero probability. This completes the proof of the strong law.

Kolmogorov's methods have had an abiding impact. Whenever the reader encounters an a.e. limit law, it is a good bet that she will find a maximal inequality and a truncation argument lurking in the background.



10 Convergence of series with random signs

The reader is well aware that the harmonic series $1 + 1/2 + 1/3 + \cdots$ is divergent. The alternating series $1 - 1/2 + 1/3 - 1/4 + \cdots$, however, is convergent (the limiting value is $\log 2$). What can one say then about the series $\sum_n \pm \frac{1}{n}$ where the signs are chosen independently, each with probability $1/2$? More generally, what is the probability that the series $\sum_n \pm c_n$ converges? Here $\{c_n\}$ represents a real sequence. The problem was posed in this form by Hugo Steinhaus in 1922. Kolmogorov's maximal inequality of the previous section provides the key to a simple resolution.

THEOREM 1 Suppose $\{X_n, n \geq 1\}$ is a sequence of independent, zero-mean random variables. Suppose the variances exist and $\sum_n \text{Var}(X_n)$ converges. Then $\sum_n X_n$ converges a.e.

PROOF: By Example XII.10.1, the set of sample points on which $\sum_n X_n$ converges is a remote event, hence has probability either zero or one. The challenge is to decide which.

As usual, consider the partial sums $S_n = X_1 + \cdots + X_n$. Fix any $\epsilon > 0$. As $S_{n+j} - S_n = X_{n+1} + \cdots + X_{n+j}$, by Kolmogorov's maximal inequality,

$$P\left(\bigcup_{1 \leq j \leq m} \{|S_{n+j} - S_n| \geq \epsilon\}\right) = P\left\{\max_{1 \leq j \leq m} |S_{n+j} - S_n| \geq \epsilon\right\} \leq \frac{1}{\epsilon^2} \sum_{j=1}^m \text{Var}(X_{n+j}).$$

The sets $\bigcup_{j=1}^m \{|S_{n+j} - S_n| \geq \epsilon\}$ form an increasing sequence of events with limit set $\bigcup_{j=1}^{\infty} \{|S_{n+j} - S_n| \geq \epsilon\}$. By proceeding to the limit as $m \rightarrow \infty$ on both sides, by continuity of probability measure, we hence obtain

$$P\left(\bigcup_{j \geq 1} \{|S_{n+j} - S_n| \geq \epsilon\}\right) = P\left\{\sup_{j \geq 1} |S_{n+j} - S_n| \geq \epsilon\right\} \leq \frac{1}{\epsilon^2} \sum_{j=1}^{\infty} \text{Var}(X_{n+j}),$$

with the bound on the right converging to zero as $n \rightarrow \infty$ as the series $\sum_n \text{Var}(X_n)$ is convergent by assumption. By Theorem XII.5.3, it hence follows that $\{S_n, n \geq 1\}$ is a Cauchy sequence, hence convergent, with probability one. ▶

EXAMPLES: 1) *The harmonic series with random signs.* The series $\sum_n \pm 1/n$ is convergent a.e. as we may identify X_n with the signed variable which takes values $\pm 1/n$ with equal probability. Then $E(X_n) = 0$ and $\text{Var}(X_n) = 1/n^2$. As $\sum_n 1/n^2$ is convergent our conclusion follows. (Pietro Mengoli proposed the problem of the evaluation of $\sum_{n \geq 1} 1/n^2$ in 1644—this is the so-called Basel problem. Leonhard Euler settled it in 1735 by showing that the series converges to $\pi^2/6$. We do not need such precision, only that the sum converges, and this is easy to see by, say, the integral test.) The convergence of the alternating harmonic series $1 - 1/2 + 1/3 - 1/4 + \dots = \log 2$ is hence typical of the general behaviour. The harmonic series $1 + 1/2 + 1/3 + 1/4 + \dots$ itself is atypical in this regard. 2) *Steinhaus's problem.* Suppose $\{c_n, n \geq 1\}$ is a sequence of real values. By identifying X_n with the signed variable $\pm c_n$ we see by a similar argument that the series $\sum_n \pm c_n$ converges a.e. if $\sum_n c_n^2$ converges.

3) *Convergence of Rademacher series.* Suppose $r_n(t)$ denotes the n th Rademacher function defined in Section V.2. By the correspondence between coin tosses and the binary digits, Steinhaus's problem is equivalent to a consideration of the series $\sum_n c_n r_n(t)$. The previous example now shows that $\sum_n c_n r_n(t)$ converges for a.e. t in the unit interval if $\sum_n c_n^2$ converges. In this form the solution was already known to Hans Rademacher at the time Steinhaus proposed the problem.¹⁴ The view from L^2 is informative.

Write $f_n(t) = \sum_{k=1}^n c_k r_k(t)$. By the orthogonality of the Rademacher functions, we have $\|f_{n+k} - f_n\|^2 = \sum_{j=n+1}^{n+k} c_j^2 \leq \sum_{j=n+1}^{\infty} c_j^2 \rightarrow 0$ as $\sum_n c_n^2$ converges. It follows that $\{f_n\}$ is a Cauchy sequence in L^2 . By Theorem XXI.3.2 in the Appendix, it follows that the series $\sum_k c_k r_k(t)$ converges also in mean-square to a limit function f .

4) *Divergence of Rademacher series.* What if $\sum_n c_n^2$ diverges? It is reasonable to guess then that $\sum_n c_n r_n(t)$ diverges a.e. Suppose $c_n \rightarrow 0$. (The answer is trivial otherwise.) By identifying binary digits with coin tosses, Kolmogorov's zero–one law says that the series $\sum_n c_n r_n(t)$ either converges a.e. or diverges a.e. To set up a contradiction, suppose hence that $\sum_n c_n r_n(t)$ converges a.e. to some measurable function $f(t)$. This function is not necessarily bounded so consider instead the function $e^{ixf(t)}$ which is bounded absolutely by 1. Here x is any non-zero real value.

For each n , write $f_n(t) = \sum_{k=1}^n c_k r_k(t)$. By the independence of the binary digits (Viète's formula (V.3.4)!), we have

$$\begin{aligned} \int_0^1 e^{ixf_n(t)} dt &= \int_0^1 \exp\left(ix \sum_{k=1}^n c_k r_k(t)\right) dt = \int_0^1 \prod_{k=1}^n \exp(ixc_k r_k(t)) dt \\ &= \prod_{k=1}^n \int_0^1 \exp(ixc_k r_k(t)) dt = \prod_{k=1}^n \cos(xc_k). \end{aligned}$$

As $c_k \rightarrow 0$, the application of two Taylor approximations, first for the cosine then the logarithm, shows that $\log \cos(xc_k) = \log(1 - x^2 c_k^2/2 + o(c_k^2)) = -x^2 c_k^2/2 + o(c_k^2)$ as $k \rightarrow \infty$. Thus, if $\sum_k c_k^2$ diverges then $\sum_{k=m}^n \log \cos(xc_k)$ diverges to $-\infty$ as $n \rightarrow \infty$. It follows that

$$\prod_{k=m}^n \cos(xc_k) = \exp\left(\sum_{k=m}^n \log(\cos(xc_k))\right) \rightarrow 0$$

¹⁴H. A. Rademacher, "Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen", *Mathematische Annalen*, vol. 87, pp. 112–138, 1922.

and so $\int_0^1 e^{ixf_n(t)} dt \rightarrow 0$ as $n \rightarrow \infty$. Now $e^{ixf_n(t)} \rightarrow e^{ixf(t)}$ for a.e. t as, by assumption, $\{f_n\}$ converges a.e. to the limit function f . By Lebesgue's dominated convergence theorem it then follows that $\int_0^1 e^{ixf_n(t)} dt \rightarrow \int_0^1 e^{ixf(t)} dt$. We hence conclude that $\int_0^1 e^{ixf(t)} dt = 0$ for every non-zero x .

On the other hand, if $\{x_n\}$ is any positive sequence decreasing to zero, say, $x_n = 1/n$, we have $x_n f(t) \rightarrow 0$, whence $e^{ix_n f(t)} \rightarrow 1$, for a.e. t . By another application of the dominated convergence theorem it follows that $\int_0^1 e^{ix_n f(t)} dt \rightarrow 1$.

The two sides of the argument lead us to the conclusion $0 = 1$ and so our hypothesis is flawed. It must hence be the case that $\sum_n c_n r_n(t)$ diverges. The reader should remark the key rôle played by independence in the argument. ►

Our findings in the examples are worth summarising.

THEOREM 2 *The series $\sum_n c_n r_n(t)$ either converges or diverges for a.e. t in the unit interval according as whether $\sum_n c_n^2$ converges or diverges.*

Proceeding along the lines of Theorem 1, Kolmogorov determined necessary and sufficient conditions for the convergence of a random series $\sum_n X_n$. This is the brutal *three series theorem*. We shall be content with Theorem 1 and not proceed further in this direction.

11 Uniform convergence per Glivenko and Cantelli

Suppose X_1, X_2, \dots, X_n is a sequence of independent random variables drawn from a common distribution F . For every real t , let $\theta_t(x) = 1_{(-\infty, t]}(x)$ be the indicator for the interval $-\infty < x \leq t$. Then the random variable $F_n(t) = F_n(t; X_1, \dots, X_n) = \frac{1}{n}(\theta_t(X_1) + \dots + \theta_t(X_n))$ connotes the *empirical distribution function* which places equal mass on each element of the random sample X_1, \dots, X_n . Now, for each j , the indicator $\theta_t(X_j)$ is a Bernoulli trial with success probability $E(\theta_t(X_j)) = P[X_j \leq t] = F(t)$ so that by the strong law of large numbers it follows that $F_n(t) \rightarrow^{\text{a.e.}} F(t)$. Thus, *the value of the d.f. $F(t)$ at any point t is asymptotically well approximated by the empirical distribution $F_n(t)$* . It is clear then that, for a given, sufficiently large, sample X_1, \dots, X_n , the values of the empirical distribution function $F_n(t_1; X_1, \dots, X_n), \dots, F_n(t_m; X_1, \dots, X_n)$ at any finite collection of points t_1, \dots, t_m are all *simultaneously* good approximations to the values of the distribution function $F(t_1), \dots, F(t_m)$ at those points: formally, for any $\epsilon > 0$,

$$P\left\{\max_{1 \leq j \leq m} |F_n(t_j) - F(t_j)| \geq \epsilon \text{ i.o.}\right\} = 0.$$

But is it true that $F_n(t; X_1, \dots, X_n)$ converges to $F(t)$ *simultaneously for all real values t* ? If this is the case then a given finite sample X_1, \dots, X_n will provide a *uniformly* good estimate of the distribution function $F(t)$ in its entirety. V. I. Glivenko and F. P. Cantelli showed independently in 1933 that this is indeed the case.¹⁵ Their theorem had unexpected ramifications—the reader will find a beguiling application in machine learning in the following section.

¹⁵V. I. Glivenko, "Sulla determinazione empirica delle leggi di probabilità", *Giorn. Ist. Ital. Attuari*, vol. 4, pp. 92–99, 1933; F. P. Cantelli, "Sulla determinazione empirica delle leggi di probabilità", *Giorn. Ist. Ital. Attuari*, vol. 4, pp. 421–424, 1933.

THE GLIVENKO–CANTELLI THEOREM *The empirical d.f. $F_n(t)$ converges uniformly to the d.f. $F(t)$ with probability one. In notation: $\sup_t |F_n(t) - F(t)| \rightarrow^{a.e.} 0$.*

The difficulty in proof is that a given finite sample has to simultaneously satisfy an infinity of convergence conditions and the aberrant sets which, by the strong law, are guaranteed to be small for each t can, in principle, accumulate to create a big aberrant set as t ranges over all real values. The finesse needed to show that bad events cannot accumulate too rapidly pulls together several themes and is well worth study.

SYMMETRISATION VIA CAUCHY

Suppose X'_1, X'_2, \dots, X'_n is a second sequence, independent of the first, and also drawn by independent sampling from the distribution F . We write $F'_n(t) = F_n(t; X'_1, \dots, X'_n) = \frac{1}{n} \sum_{j=1}^n \theta_t(X'_j)$ for the empirical distribution of the second sample X'_1, \dots, X'_n . If F_n converges a.e. to F then so does F'_n (as the two samples have the same distribution) in which case F_n and F'_n must be close. The same theme is echoed in Theorem XIV.10.2.

LEMMA 1 (SYMMETRISATION INEQUALITY) *Fix any $\epsilon > 0$. If $n \geq 2/\epsilon^2$ then¹⁶*

$$P\{\sup_t |F_n(t) - F(t)| \geq \epsilon\} \leq 2 P\{\sup_t |F_n(t) - F'_n(t)| \geq \epsilon/2\}.$$

PROOF: Vector notation helps simplify presentation. Write $x = (x_1, \dots, x_n)$ and $X = (X_1, \dots, X_n)$, with similar notations also for the primed variables. Then (X, X') constitutes a double sample drawn from the product measure $F^{\otimes n} \otimes F^{\otimes n} = F^{\otimes 2n}$.

Let \mathbb{A} denote the collection of points x in \mathbb{R}^n on which $\sup_t |F_n(t; x) - F(t)| \geq \epsilon$. If the first sample X is in \mathbb{A} then there exists a real number $\tau = \tau(X)$ such that $|F_n(\tau) - F(\tau)| \geq \epsilon$. Thus, by the triangle inequality, if $X \in \mathbb{A}$, we have

$$\begin{aligned} \sup_t |F_n(t) - F'_n(t)| &\geq |F_n(\tau) - F'_n(\tau)| \\ &\geq |F_n(\tau) - F(\tau)| - |F'_n(\tau) - F(\tau)| \geq \epsilon - |F'_n(\tau) - F(\tau)| \end{aligned}$$

for all realisations of the second sample $X' \in \mathbb{R}^n$. We introduce the random variable $\rho_n = \rho_n(X, X') = \sup_t |F_n(t) - F'_n(t)|$ with the dependence on the double sample (X, X') implicit. Then, on the set of sample points for which $X \in \mathbb{A}$, we have $\rho_n \geq \epsilon/2$ if $|F'_n(\tau) - F(\tau)| \leq \epsilon/2$. Since τ is determined by X it is independent of X' and acts as a fixed index value if $X = x$ is fixed. By Fubini's theorem it follows that

$$P\{\rho_n \geq \epsilon/2\} \geq \int_{\mathbb{A}} P\{|F'_n(\tau(x)) - F(\tau(x))| \leq \epsilon/2\} dF^{\otimes n}(x). \quad (11.1)$$

Now, for every t , by Chebyshev's inequality,

$$P\{|F'_n(t) - F(t)| \leq \epsilon/2\} \geq 1 - \frac{4F(t)(1 - F(t))}{n\epsilon^2} \geq 1 - \frac{1}{n\epsilon^2}$$

¹⁶The reader who is worried about measurability as the supremum is taken over the uncountable set of the reals may restrict the supremum to the countable set of the rationals—measurability is now taken care of by Theorem XII.5.1—and, as the rationals form a countably dense subset of the real line, proceed to the entire line by limiting arguments exploiting the right continuity of the d.f.

as $4x(1-x) \leq 1$ for all x . The right-hand side is at least $1/2$ if $n \geq 2/\epsilon^2$ and *a fortiori* the integrand on the right in (11.1) is uniformly bounded below by $1/2$. It follows that $\mathbf{P}\{\rho_n \geq \epsilon/2\} \geq \mathbf{P}\{\mathbf{X} \in \mathbb{A}\}/2$ if $n \geq 2/\epsilon^2$ as was to be shown. ▶

If we identify $X_{n+j} = X'_j$, then $F_n(t) - F_{2n}(t) = \frac{1}{2}(F_n(t) - F'_n(t))$ and the symmetrisation inequality says that if F_n is uniformly close to F_{2n} then F_n converges uniformly. The reader will recognise that this is the essential idea behind a Cauchy sequence: to show that a real sequence $\{x_n\}$ converges it is sufficient to show that it is Cauchy, that is to say, $\sup_k |x_n - x_{n+k}| \rightarrow 0$ as $n \rightarrow \infty$. Cast in this light our lemma codifies a probabilistic Cauchy principle.

The value of the symmetrisation lemma lies in the observation that a comparison of an empirical distribution with the true but unknown distribution has been replaced by a comparison of two empirical distributions (which are known once the double sample is specified). A delicate further symmetrisation now helps eliminate the notational nuisance of the double sample and reduces consideration to a random walk.

SYMMETRISATION VIA AUXILIARY RANDOMISATION

By collecting terms, we may write $F_n(t) - F'_n(t) = \frac{1}{n} \sum_{j=1}^n \xi_t(X_j, X'_j)$ where the summands $\xi_t(X_j, X'_j) = \theta_t(X_j) - \theta_t(X'_j)$ are obtained by symmetrisation of the Bernoulli distribution,

$$\xi_t(X_j, X'_j) = \begin{cases} -1 & \text{with probability } F(t)(1 - F(t)), \\ 0 & \text{with probability } 1 - 2F(t)(1 - F(t)), \\ +1 & \text{with probability } F(t)(1 - F(t)). \end{cases}$$

If we interchange the order of X_j and X'_j we see that $\xi_t(X'_j, X_j)$ simply reverses the sign of $\xi_t(X_j, X'_j)$ and, in view of the symmetry of the distribution, leaves the distribution unchanged. As the pairs $\{(X_j, X'_j), 1 \leq j \leq n\}$ are independent, so too are the variables $\{\xi_t(X_j, X'_j), 1 \leq j \leq n\}$ and so the distribution of $\rho_n = \sup_t \frac{1}{n} |\sum_{j=1}^n \xi_t(X_j, X'_j)|$ is invariant to pairwise exchanges of the variables X_j and X'_j . Let $\Pi = (Y_1 Y_2 \dots Y_n) (Y'_1 Y'_2 \dots Y'_n)$ be any double sequence obtained by pairwise exchanges of elements from the original double sequence $(X_1 X_2 \dots X_n) (X'_1 X'_2 \dots X'_n)$ and let $\rho_n(\Pi) = \sup_t \frac{1}{n} |\sum_{j=1}^n \xi_t(Y_j, Y'_j)|$ be the corresponding value of ρ_n for the permuted double sequence. Then $\rho_n(\Pi)$ has the same distribution as ρ_n for every choice of permutation Π . We have no reason to prefer any one permutation of the 2^n possible permutations obtained by pairwise exchanges of elements in the double sample and so we symmetrise by selecting a random permutation Π . We accordingly have $\mathbf{P}\{\rho_n \geq \epsilon/2\} = \mathbf{P}\{\rho_n(\Pi) \geq \epsilon/2\}$, the probability on the right averaging not only over the double sample but also over the choice of permutation.

A little notation will help clarify why this is progress. In a random permutation Π , each pair (X_j, X'_j) is interchanged to form the pair (X'_j, X_j) with probability one-half, these interchanges occurring independently across j . An interchange of the j th coordinates means a change in sign of $\xi_t(X_j, X'_j)$ to $\xi_t(X'_j, X_j) = -\xi_t(X_j, X'_j)$. Accordingly, if $\{\sigma_j, 1 \leq j \leq n\}$ is an ancillary sequence of independent random variables, each σ_j taking values -1 and $+1$ only, each with equal probability one-half, then we may write

$$\rho_n(\Pi) = \frac{1}{n} \sum_{j=1}^n \sigma_j \xi_t(X_j, X'_j) = \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X_j) - \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X'_j),$$

whence, by the triangle inequality, $|\rho_n(\Pi)| \leq \left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X_j) \right| + \left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X'_j) \right|$. In consequence, the occurrence of the event $|\rho_n(\Pi)| \geq \epsilon/2$ implies the occurrence of at least one of the events $\left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X_j) \right| \geq \epsilon/4$ or $\left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X'_j) \right| \geq \epsilon/4$. By Boole's inequality, it follows that

$$P\left\{\rho_n(\Pi) \geq \frac{\epsilon}{2}\right\} \leq P\left\{\sup_t \left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X_j) \right| \geq \frac{\epsilon}{4}\right\} + P\left\{\sup_t \left| \frac{1}{n} \sum_{j=1}^n \sigma_j \theta_t(X'_j) \right| \geq \frac{\epsilon}{4}\right\}.$$

The two probabilities on the right are equal as the sequences $\{X_j, 1 \leq j \leq n\}$ and $\{X'_j, 1 \leq j \leq n\}$ have the same distribution and it suffices hence to consider the sums $R_n(t) = R_n(t; X) = \sum_{j=1}^n \sigma_j \theta_t(X_j)$ representing symmetric random walks about the origin. Summarising, we have

LEMMA 2 *For every $\epsilon > 0$, we have $P\{\rho_n \geq \epsilon/2\} \leq 2 P\{\sup_t |R_n(t)| \geq \epsilon n/4\}$.*

THE VALUE OF AN EXPONENTIAL BOUND

We've now reduced considerations to the behaviour in the tails of the partial sums of indicators with random signs. There is still work to be done, however, as a supremum over real values has yet to be dealt with.

The evaluation of $P\{\sup_t |R_n(t)| \geq \epsilon n/4\}$ requires a joint integration over the variables $X = (X_1, \dots, X_n)$ and $\sigma = (\sigma_1, \dots, \sigma_n)$. Fubini's theorem helps mop up. Iterating the integrals in the indicated order shows that

$$P\{\sup_t |R_n(t; X)| \geq \epsilon/4\} = \int_{\mathbb{R}^n} P\{\sup_t |R_n(t; x)| \geq \epsilon n/4\} dF^{\otimes n}(x). \quad (11.2)$$

For any $x \in \mathbb{R}^n$, the expression $R_n(t; x) = \sum_{j=1}^n \sigma_j \theta_t(x_j) = \sum \pm 1$ sums over $m \leq n$ sign variables with only those j (say, $m = m(x)$ in number) for which $\theta_t(x_j) = 1$ contributing to the sum. Thus, as t ranges over all real values with x fixed, the sums $R_n(t; x)$ vary over only the $m+1$ distinct values, $\pm m, \pm(m-2), \dots$. Suppose t_0, t_1, \dots, t_m are representative values of t for which the sum $R_n(t; x)$ takes distinct values. Then, for any given $x \in \mathbb{R}^n$, Boole's inequality coupled with Hoeffding's inequality (1.1') for a simple random walk shows that

$$\begin{aligned} P\{\sup_t |R_n(t; x)| \geq \epsilon n/4\} &= P\{\max_{0 \leq j \leq m} |R_n(t_j; x)| \geq \epsilon n/4\} \\ &\leq \sum_{j=0}^m P\{|R_n(t_j; x)| \geq \epsilon n/4\} \leq 2(m+1)e^{-n\epsilon^2/32} \leq 2(n+1)e^{-n\epsilon^2/32}, \end{aligned} \quad (11.3)$$

the bound on the right uniform in x . Substituting for the integrand on the right in (11.2) yields an exponentially decaying bound for the probability on the left.

LEMMA 3 *For every $\epsilon > 0$, we have $P\{\sup_t |R_n(t)| \geq \epsilon n/4\} \leq 2(n+1)e^{-n\epsilon^2/32}$.*

All that is needed now is to stitch the conclusions of the lemmas together.

THEOREM 2 *The sequence of empirical distributions $\{F_n(t), n \geq 1\}$ converges in probability to the d.f. $F(t)$ uniformly in t . More precisely, suppose $\epsilon > 0$ and $n \geq 2/\epsilon^2$. Then*

$$P\{\sup_t |F_n(t) - F(t)| \geq \epsilon\} \leq 8(n+1)e^{-n\epsilon^2/32}. \quad (11.4)$$

The upper bound dies quickly, so quickly indeed that $\sum_n P\{\sup_t |F_n(t) - F(t)| \geq \epsilon\}$ converges, the exponential decay rapidly quenching the linear growth in the summand. The almost everywhere convergence asserted by the Glivenko–Cantelli theorem is now taken care of by the Borel–Cantelli lemma and the proof is complete. The value of the exponential bound is evident here. Chebyshev's inequality would only give an upper bound of $16(m+1)m/(\epsilon^2 n^2)$ in (11.3) which is much too large to be useful.

12 What can be learnt per Vapnik and Chervonenkis

Thirty five years after the publication of the Glivenko–Cantelli theorem, V. N. Vapnik and A. Ya. Chervonenkis produced a profound generalisation.¹⁷

VAPNIK–CHERVONENKIS CLASSES

Restating the Glivenko–Cantelli theorem in terms of set functions suggests a profitable direction for inquiry. Writing $\mathfrak{R} = \{(-\infty, t] : t \in \mathbb{R}\}$ for the family of rays, the Glivenko–Cantelli theorem says that $\sup_{A \in \mathfrak{R}} |F_n(A) - F(A)| \rightarrow^{a.e.} 0$. If we replace the family \mathfrak{R} of rays by a family \mathfrak{A} of measurable subsets, then, reusing notation and writing $\theta_A(X) = 1\{X \in A\}$ for any $A \in \mathfrak{A}$, the strong law says that, for any given event $\{X \in A\}$, the empirical frequency of its occurrence $F_n(A) = \frac{1}{n}(\theta_A(X_1) + \dots + \theta_A(X_n))$ converges a.e. to the event probability $F(A)$. The issue at hand is whether there is *uniform* a.e. convergence over the elements of \mathfrak{A} . A close examination of the proof of the Glivenko–Cantelli theorem shows that the symmetrisation steps are unaffected¹⁸ and that the only place where we used the *topological* structure of the rays $(-\infty, t]$ was in the final step, in the reduction of a supremum over t to a finite collection of no more than $n+1$ events. In a slightly more vivid language, the key to the proof was the observation that the rays $(-\infty, t]$ pick out no more than $n+1$ distinct subsets of the sample $\{X_1, \dots, X_n\}$ as t ranges over all real values. Some new notation and terminology helps bring the essential element into focus.

We begin with an abstract probability space¹⁹ $(\mathcal{X}, \mathcal{F}, F)$ and suppose \mathfrak{A} is any family of measurable subsets of \mathcal{X} . For each n , we deal with the product space \mathcal{X}^n equipped with product measure $P = F^{\otimes n}$. Given any sample $x = (x_1, \dots, x_n)$ in \mathcal{X}^n , each set $A \in \mathfrak{A}$ picks out a subsample $x^A = (x_j, j \in J^A)$ where J^A is the subset of indices j for which $x_j \in A$. As A varies across the family \mathfrak{A} , the distinct subsamples of x that are picked out form a family $\Delta_{\mathfrak{A}}(x)$ of subsamples called the *trace* of \mathfrak{A} on x . As there are 2^n

¹⁷For an English translation of their 1968 work see V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities", *Theory of Probability and Its Applications*, vol. 16, pp. 264–280, 1971.

¹⁸In the second symmetrisation step Vapnik and Chervonenkis symmetrise by permutation of the *entire* double sample and obtain exponential bounds by a consideration of the tails of the hypergeometric distribution; see Problems 29–31. I learnt of the possibilities of symmetrisation by pairwise exchanges from D. Pollard, *Convergence of Stochastic Processes*. New York: Springer-Verlag, 1984. This method of auxiliary randomisation simplifies the proof at the expense of slightly worsened constants and exponents.

¹⁹Finesse measurability difficulties by assuming \mathcal{X} is a Polish space, that is to say, a metric space containing a countably dense subset of points and in which every Cauchy sequence converges.

possible subsamples of $x = (x_1, \dots, x_n)$ [including the empty subsample and the entire sample (x_1, \dots, x_n) itself] it is clear that $1 \leq \text{card } \Delta_{\mathfrak{A}}(x_1, \dots, x_n) \leq 2^n$.

Some picturesque terminology has taken root in this context: we say that the sample x is *shattered* by \mathfrak{A} if $\text{card } \Delta_{\mathfrak{A}}(x) = 2^n$; that is to say, \mathfrak{A} picks out every subsample of $x = (x_1, \dots, x_n)$. The terminology is a little fervid as it invites the reader to imagine an aggressive \mathfrak{A} breaking (x_1, \dots, x_n) into minute pieces instead of the reality of an assiduous \mathfrak{A} picking up all the subsamples of (x_1, \dots, x_n) , large and small, but the language at least has the virtue of being vivid.

The *growth function* $D_{\mathfrak{A}}(n) = \max\{\text{card } \Delta_{\mathfrak{A}}(x) : x \in \mathcal{X}^n\}$ is the size of the largest trace as we run through the points in \mathcal{X}^n . Thus, $1 \leq D_{\mathfrak{A}}(n) \leq 2^n$. The family \mathfrak{A} becomes interesting if the growth rate is not too large.

DEFINITION 1 We say that a family \mathfrak{A} of measurable subsets of a (Polish) space \mathcal{X} forms a *Vapnik–Chervonenkis class* if the growth function $D_{\mathfrak{A}}(n)$ is dominated by a polynomial in n . In a more picturesque language, a Vapnik–Chervonenkis class \mathfrak{A} picks out only a polynomial number of subsets of any given collection of points.

The family of rays $\mathfrak{R} = \{(-\infty, t] : t \in \mathbb{R}\}$ is a Vapnik–Chervonenkis class. Indeed, $D_{\mathfrak{R}}(n) = n + 1$ as exactly $n + 1$ subsets of any distinct collection of n real values are picked out by \mathfrak{R} and so the growth function has only a linear growth with n .

As n becomes large a Vapnik–Chervonenkis class \mathfrak{A} picks out an asymptotically small number $D_{\mathfrak{A}}(n)$ of the 2^n subsets of any given set of n elements. This was key in the proof of the Glivenko–Cantelli theorem. If we replace the family \mathfrak{R} of rays by a general Vapnik–Chervonenkis class \mathfrak{A} then the proof of the Glivenko–Cantelli theorem carries through almost verbatim with only the replacement of scattered references to rays by sets $A \in \mathfrak{A}$ and the replacement of the factor $n + 1$ in the final step by the polynomially dominated growth function $D_{\mathfrak{A}}(n)$. We hence conclude that

$$P\left\{\sup_{A \in \mathfrak{A}} |F_n(A) - F(A)| \geq \epsilon\right\} \leq 8 E(\text{card } \Delta_{\mathfrak{A}}(X)) e^{-n\epsilon^2/32} \leq 8 D_{\mathfrak{A}}(n) e^{-n\epsilon^2/32}. \quad (12.1)$$

Uniform a.e. convergence falls out via the Borel–Cantelli lemma as before because the exponential dominates the polynomial increase of the growth function.

THE VAPNIK–CHERVENENKIS THEOREM Suppose \mathfrak{A} is a Vapnik–Chervonenkis class, F any sampling distribution. Then, with probability one, the empirical frequencies $F_n(A)$ converge to the event probabilities $F(A)$ uniformly in \mathfrak{A} . In notation: $\sup_{A \in \mathfrak{A}} |F_n(A) - F(A)| \rightarrow 0$ a.e.

The question that arises naturally now is whether there is a simple characterisation of Vapnik–Chervonenkis classes. In their elegant paper Vapnik and Chervonenkis showed that such classes of sets have an essentially geometric character.

THE GEOMETRY OF THE SITUATION

The family \mathfrak{R} of rays will pick out all subsamples of any singleton (x) , namely the empty subsample and the singleton sample (x) itself, but will pick out only three of the four possible subsamples of a pair of points (x_1, x_2) . Intuition suggests along these lines that, if the class \mathfrak{A} is sufficiently rich, then \mathfrak{A} will be able to pick out all subsamples of a given sample (x_1, \dots, x_n) if n is small (and provided that the sample is not pathological). If \mathfrak{A}

is a Vapnik–Chervonenkis class, however, then, as n increases, sooner or later \mathcal{A} will fail to pick out all the subsamples of *any* sample of n points. Thus, the *largest* value of n for which there exists a sample (x_1, \dots, x_n) which is shattered by \mathcal{A} becomes of interest.

DEFINITION 2 The *Vapnik–Chervonenkis dimension of the class \mathcal{A} with respect to the sample $x = (x_1, \dots, x_n)$* , which we denote $V_{\mathcal{A}}(x)$, is the size of the largest subsample of x that is shattered by \mathcal{A} . The *Vapnik–Chervonenkis dimension of the class \mathcal{A}* , which we denote $V_{\mathcal{A}}$, is the size of the largest sample that is shattered by \mathcal{A} . Thus, $V_{\mathcal{A}} = \sup V_{\mathcal{A}}(x_1, \dots, x_n)$ where the supremum is over all n and all n -samples (x_1, \dots, x_n) . If \mathcal{A} does not shatter unboundedly large samples then $V_{\mathcal{A}}$ as the largest integer n for which $D_{\mathcal{A}}(n) = 2^n$.

Remarkably, this apparently crude, geometric idea of dimensionality is key to the characterisation of the growth function.

THEOREM 2 Suppose \mathcal{A} has finite Vapnik–Chervonenkis dimension $V = V_{\mathcal{A}}$. Then the growth function is bounded by $D_{\mathcal{A}}(n) \leq \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{V}$ and, *a fortiori*, $D_{\mathcal{A}}(n) \leq 1 + n^V$.

PROOF: A beautiful induction argument lies at the heart of the proof. Write $B_V(m) = \binom{m}{0} + \binom{m}{1} + \dots + \binom{m}{V}$. We shall show by induction that $D_{\mathcal{A}}(n) \leq B_{V_{\mathcal{A}}}(n)$. The base of the induction is trite as $D_{\mathcal{A}}(n) = 1$ if $n = 0$ or $V_{\mathcal{A}} = 0$. As induction hypothesis, suppose now that, for a family of sets \mathcal{B} in an arbitrary space, we have $D_{\mathcal{B}}(m) \leq B_V(m)$ whenever $m \leq n$ and $V_{\mathcal{B}} \leq V$. To set up the induction suppose \mathcal{A} is a family of sets with $V_{\mathcal{A}} = V$, $x = (x_1, \dots, x_n)$ is any n -sample, and $\Delta_{\mathcal{A}}(x)$ is the trace of \mathcal{A} on x . We consider the effect of increasing dimensionality by expanding x by addition of a new point x_{n+1} . The subsamples $t \in \Delta_{\mathcal{A}}(x_1, \dots, x_n)$ are of two types: say that t is of the first type if either $t \in \Delta_{\mathcal{A}}(x_1, \dots, x_n, x_{n+1})$ or $(t, x_{n+1}) \in \Delta_{\mathcal{A}}(x_1, \dots, x_n, x_{n+1})$ [but not both]; say that t is of the second type if both subsamples t and (t, x_{n+1}) of the expanded sample (x, x_{n+1}) are in the new trace $\Delta_{\mathcal{A}}(x_1, \dots, x_n, x_{n+1})$. Then \mathcal{A} is partitioned into two disjoint subfamilies of sets, those sets \mathcal{A}' which pick out subsamples of the first type forming a subclass \mathcal{A}' and those sets \mathcal{A}'' which pick out subsamples of the second type forming a disjoint subclass \mathcal{A}'' . Suppose there are $I = \text{card } \Delta_{\mathcal{A}'}(x)$ subsamples of the first type and $J = \text{card } \Delta_{\mathcal{A}''}(x)$ subsamples of the second type in the trace $\Delta_{\mathcal{A}}(x_1, \dots, x_n)$. Then $\text{card } \Delta_{\mathcal{A}}(x_1, \dots, x_n) = I + J$.

With the addition of the point x_{n+1} to the sample (x_1, \dots, x_n) , each subsample t of the first type in the old trace $\Delta_{\mathcal{A}}(x_1, \dots, x_n)$ engenders a unique subsample, either t or (t, x_{n+1}) , in the new trace $\Delta_{\mathcal{A}}(x_1, \dots, x_n, x_{n+1})$, while each subsample t of the second type in the old trace engenders two distinct subsamples, both t and (t, x_{n+1}) , in the new trace. Thus,

$$\text{card } \Delta_{\mathcal{A}}(x_1, \dots, x_n, x_{n+1}) = I + 2J = \text{card } \Delta_{\mathcal{A}}(x_1, \dots, x_n) + \text{card } \Delta_{\mathcal{A}''}(x_1, \dots, x_n). \quad (12.2)$$

To estimate the number of subsamples of the second type in the old trace we observe that \mathcal{A}'' cannot shatter any subsample of x of size V . Indeed, if t is a subsample of the second type that is shattered by \mathcal{A}'' then the expanded subsample (t, x_{n+1}) is also shattered by \mathcal{A}'' and, *a fortiori*, also by \mathcal{A} . If t has size V this would imply that \mathcal{A} shatters a sample of size $V + 1$. Contradiction. It follows that $V_{\mathcal{A}''}(x_1, \dots, x_n) \leq V - 1$.

All is set for the induction step. The traces on the right in (12.2) deal with the restriction of the families \mathcal{A} and \mathcal{A}'' to the finite set $\{x_1, \dots, x_n\}$ which now serves in

the rôle of the universal space. Accordingly, write \mathfrak{A}_r and \mathfrak{A}_r'' for the families of sets obtained by intersecting the elements of \mathfrak{A} and \mathfrak{A}'' , respectively, with $\{x_1, \dots, x_n\}$. Then $V_{\mathfrak{A}_r} = V_{\mathfrak{A}}(x_1, \dots, x_n) \leq V_{\mathfrak{A}} = \nu$, by assumption, and, as we have just seen, $V_{\mathfrak{A}_r''} = V_{\mathfrak{A}''}(x_1, \dots, x_n) \leq \nu - 1$. Thus, by two applications of the induction hypothesis, we have $\text{card } \Delta_{\mathfrak{A}}(x_1, \dots, x_n) = D_{\mathfrak{A}_r}(n) \leq B_{\nu}(n)$ and $\text{card } \Delta_{\mathfrak{A}''}(x_1, \dots, x_n) = D_{\mathfrak{A}_r''}(n) \leq B_{\nu-1}(n)$. It follows that $\text{card } \Delta_{\mathfrak{A}}(x_1, \dots, x_n, x_{n+1}) \leq B_{\nu}(n) + B_{\nu-1}(n)$. The sum on the right collapses by repeated applications of Pascal's triangle as, by grouping terms,

$$\begin{aligned} B_{\nu}(n) + B_{\nu-1}(n) &= \binom{n}{0} + \{\binom{n}{1} + \binom{n}{0}\} + \{\binom{n}{2} + \binom{n}{1}\} + \cdots + \{\binom{n}{\nu} + \binom{n}{\nu-1}\} \\ &= \binom{n+1}{0} + \binom{n+1}{1} + \binom{n+1}{2} + \cdots + \binom{n+1}{\nu} = B_{\nu}(n+1). \end{aligned} \quad (12.3)$$

Thus, $\text{card } \Delta_{\mathfrak{A}}(x_1, \dots, x_n, x_{n+1}) \leq B_{\nu}(n+1)$ for any choice of sample $(x_1, \dots, x_n, x_{n+1})$ and so $D_{\mathfrak{A}}(n+1) \leq B_{\nu}(n+1)$. This completes the induction.

To finish off the proof we have to show that $B_{\nu}(n) \leq 1 + n^{\nu}$ for positive ν and n . The obvious boundary cases $B_0(n) = B_{\nu}(0) = 1$ set up another inductive verification. If ν and n are strictly positive, then

$$B_{\nu}(n+1) \stackrel{(i)}{=} B_{\nu}(n) + B_{\nu-1}(n) \stackrel{(ii)}{\leq} (1 + n^{\nu}) + (1 + n^{\nu-1}) \stackrel{(iii)}{\leq} 1 + (n+1)^{\nu},$$

where (i) is a restatement of the Pascal recurrence (12.3), (ii) follows by induction hypothesis, and (iii) follows by the binomial theorem. ▶

EXAMPLES: 1) *Rays, again.* If $\mathcal{X} = \mathbb{R}$ and \mathfrak{R} is the family of rays then $V_{\mathfrak{R}} = 1$ and $D_{\mathfrak{R}}(n) = 1 + n$.

2) *Open sets.* If $\mathcal{X} = (0, 1)$ and \mathfrak{D} consists of the open sets then $V_{\mathfrak{D}} = \infty$ and $D_{\mathfrak{D}}(n) = 2^n$ for each n .

3) *Closed discs.* If $\mathcal{X} = \mathbb{R}^2$ and \mathfrak{D} consists of all closed discs in the plane then $V_{\mathfrak{D}} = 3$.

4) *Half-spaces.* Suppose $\mathcal{X} = \mathbb{R}^{\nu}$ and $\mathfrak{H} = \{ \mathbb{H}_{\mathbf{w}}, \mathbf{w} \in \mathbb{R}^{\nu} \}$ is the set of all half-spaces of the form $\mathbf{x}\mathbf{w}^T \geq 0$ for each fixed vector $\mathbf{w} \in \mathbb{R}^{\nu}$. Then $V_{\mathfrak{H}} = \nu$. ▶

The representative examples I have included above range from easy to sophisticated. I will leave these (with hints for solution where the logic is slippery) together with a selection of others for the reader to attempt in the *Problems*. Problem 28 shows how one can bootstrap examples of this kind to create a huge range of families of finite Vapnik–Chervonenkis dimension, for all of whom the Vapnik–Chervonenkis theorem of uniform convergence holds sway.

A QUESTION OF IDENTIFICATION

The Glivenko–Cantelli theorem provides theoretical cover for the empirical estimation of distributions. By reversing the question it may also be put to use to answer a question on the complexity of identification. Here is the setting.

A ray $(-\infty, t_0]$ represents an underlying unknown state of nature to be identified. To this end a supplicant goes to an oracle who provides a random sample X_1, \dots, X_n obtained by independent sampling from a distribution F that is concealed, together with a set of labels $Y_1 = \theta_{t_0}(X_1), \dots, Y_n = \theta_{t_0}(X_n)$ which identify whether a given point X_j lies in the underlying interval $(-\infty, t_0]$ or not. The goal now is to identify the ray $(-\infty, t_0]$, at least approximately, by use of the *labelled sample* $\{(X_j, Y_j), 1 \leq j \leq n\}$.

Naturally enough, the acolyte picks as a hypothesis any ray $(-\infty, \tau]$ which agrees with the labels of the given sample, $\theta_\tau(X_j) = Y_j$ for each j . Thus, for instance, if a is the largest value of the X_j for which $Y_j = 1$ and b is the smallest value of the X_j for which $Y_j = 0$, we may select for the right endpoint any value τ in the interval $[a, b]$. How good is such an estimate τ of the unknown t_0 ?

As a measure of goodness we ask how likely it is that a random test element X , again provided by the oracle by independent sampling from the same, concealed distribution F , is accurately labelled by the test hypothesis $(-\infty, \tau]$. An error is made if X happens to fall between τ and t_0 and this has probability $F((-\infty, t_0] \Delta (-\infty, \tau]) = |F(\tau) - F(t_0)|$. The error probability itself is a random variable as τ is determined by the random sample. As we need a guarantee that the estimate is good whatever the underlying ray $(-\infty, t_0]$ provided by nature, we are hence led to consider the probability $P\{\sup_{t_0} |F(\tau) - F(t_0)| < \epsilon\}$ that the error is uniformly small for any selection of t_0 . Now, by two applications of the triangle inequality,

$$|F(\tau) - F(t_0)| \leq |F(\tau) - F_n(\tau)| + |F_n(\tau) - F_n(t_0)| + |F_n(t_0) - F(t_0)|.$$

The central term on the right is identically zero as the hypothesis ray $(-\infty, \tau]$ is chosen to agree with the labels of the given sample. It follows that the occurrence of the bad event $\sup_{t_0} |F(\tau) - F(t_0)| \geq \epsilon$ implies the occurrence of at least one of $\sup_{t_0} |F(\tau) - F_n(\tau)| \geq \epsilon/2$ or $\sup_{t_0} |F_n(t_0) - F(t_0)| \geq \epsilon/2$. By Boole's inequality and the Glivenko–Cantelli bound (11.4) it follows that, for every $\epsilon > 0$ and $\delta > 0$,

$$P\{\sup_{t_0} |F(\tau) - F(t_0)| \geq \epsilon\} \leq 16(n+1)e^{-n\epsilon^2/128} \leq \delta,$$

eventually. Thus, if the sample size $n = n(\epsilon, \delta)$ is sufficiently large the hypothesis $(-\infty, \tau]$ obtained from the random sample identifies the underlying ray $(-\infty, t_0]$ with error no more than ϵ and confidence at least $1 - \delta$ for any specification of t_0 and any selection of sampling distribution F . While we have made no extraordinary attempts to rein in our conservative bounds it is clear that, for a given error tolerance ϵ and a desired confidence $1 - \delta$, the requisite sample size $n = n(\epsilon, \delta)$ grows only polynomially in $1/\epsilon$ and logarithmically in $1/\delta$. A numerical estimate may carry more conviction for the reader: a sample size of 3.2×10^7 ensures identification of any ray with an error of less than 1% and a confidence in excess of 99% for any sampling distribution.

The analysis carries over without essential change to a machine learning setting where a learner wishes to identify an underlying hidden set (or *concept*) \mathbb{A}_0 in a Vapnik–Chervonenkis class \mathfrak{A} of sets given a random labelled sample $\{(X_j, Y_j), 1 \leq j \leq n\}$ provided by an oracle. Here $Y_j = \theta_{\mathbb{A}_0}(X_j) = 1\{X_j \in \mathbb{A}_0\}$. The Vapnik–Chervonenkis theorem then says that a sufficiently large finite sample of size n determined solely by the Vapnik–Chervonenkis dimension $V_{\mathfrak{A}}$, the admissible error margin ϵ , and the desired confidence $1 - \delta$, will suffice to identify \mathbb{A}_0 with error no more than ϵ and confidence at least $1 - \delta$ whatever the selection of $\mathbb{A}_0 \in \mathfrak{A}$ and the underlying distribution F . A pithy slogan captures the salient features of the result. (The picky reader who prefers a formula to an exhortation will find things fleshed out in Problems 29–31 where she will also find the germs of a more careful analysis spelled out.)

SLOGAN *In principle, concepts in a Vapnik–Chervonenkis class can be identified, with small error and high confidence, by a finite sample whose size is invariant with respect to the concept to be identified and the underlying probability law.*

This is important in typical machine learning settings where it is not only the underlying state of nature that is not known but the nature of the sampling distribution itself is also unknown.

The description of the learning process that I have provided (sans a consideration of learning time complexity) essentially lays out the idea behind probably approximately correct (or PAC) learning which was developed by L. Valiant in 1984 as a framework for the analysis of machine learning.²⁰ In 1989, A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth were apparently the first to appreciate the utility of the Vapnik–Chervonenkis formulation in Valiant’s setting.²¹ An immense literature on the subject has since burgeoned.

13 Problems

Chebyshev’s inequality, the law of large numbers, and ramifications:

1. Suppose $\text{Var } X = 0$. Show that X is equal to a constant a.e.

2. *The tails of the hypergeometric distribution.* Suppose that there are m aberrant members of a large population of $2n$ individuals. Suppose the population is split randomly into two subpopulations of size n . How likely is it that the two subpopulations will agree on the degree of aberration? If N and N' , respectively, denote the number of aberrant members of each of the two subpopulations, then $v = N/n$ and $v' = N'/n$ denote the relative frequency of aberrant individuals in these populations. Suppose $p = m/2n$ is fixed and n is large. Show that $P\{|v - p| \geq \epsilon/2\} \rightarrow 0$ and hence that $P\{|v - v'| \geq \epsilon\} \rightarrow 0$ as $n \rightarrow \infty$.

3. *Continuation, an exponential tail bound for the hypergeometric distribution.* Improve the result of the previous problem by showing that $P\{|v - v'| \geq \epsilon\} \leq 2e^{-\epsilon^2 n/2}$. [See also Problem VIII.28.]

4. *Importance sampling.* We wish to evaluate $I = \int_0^1 g(x) dx$ by sampling from a density f with support in the unit interval. Let $J = \frac{1}{n} \left(\frac{g(X_1)}{f(X_1)} + \dots + \frac{g(X_n)}{f(X_n)} \right)$. Show that $J \xrightarrow{a.e.} I$. Argue that the estimator obtained by sampling according to the *importance density* f instead of the uniform distribution is more efficient if f is chosen so that $\text{Var}(J)$ is much smaller than $\int_0^1 (g(x) - I)^2 dx$. Suppose a is a large positive constant and $g(x) = a$ if $1/3 < x < 2/3$ and $g(x) = 2a$ otherwise. By comparing variances, design an importance density f for the function g which gives better performance than the uniform density.

5. K_4 . Determine the critical rate of decay for the edge probability $p = p_n$ in the random graph $G_{n,p}$ at which a threshold function or phase transition is manifest for the property that $G_{n,p}$ contains a K_4 , a complete subgraph on four vertices.

6. *Approximating derivatives.* Suppose f has a continuous derivative in $[0, 1]$. Then the derivatives f'_n of the Bernstein polynomials (5.3) converge uniformly to f' .

²⁰L. Valiant, “A theory of the learnable”, *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

²¹A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, “Learnability and the Vapnik–Chervonenkis dimension”, *Journal of the ACM*, vol. 36, no. 4, pp. 929–965, 1989.

7. *Bernstein polynomials in two dimensions.* If $f(x, y)$ is continuous in the triangle $x \geq 0, y \geq 0, x + y \leq 1$, then $f_n(x, y) = \sum f\left(\frac{j}{n}, \frac{k}{n}\right) \frac{n!}{j!k!(n-j-k)!} x^j y^k (1-x-y)^{n-j-k}$ converges uniformly to $f(x, y)$.

8. *Generalising Bernstein's argument.* For each t in a finite or infinite interval suppose $\{F_{n,t}, n \geq 1\}$ is a sequence of distributions with mean t and variance $\sigma_{n,t}^2$. If f is continuous and bounded write $f_n(t) = \int f(x) dF_{n,t}(x)$ for the expectation of f with respect to $F_{n,t}$. By mimicking Bernstein's proof of Weierstrass's theorem, prove the following generalisation: *If $\sigma_{n,t}^2 \rightarrow 0$ then $f_n(t) \rightarrow f(t)$, the convergence being uniform in any finite interval on which $\sigma_{n,t}^2 \rightarrow 0$ uniformly.*

9. *Continuation, approximations of a different character.* Suppose f is continuous and bounded on $(0, \infty)$. For each n , identify $f_n(t)$ in turn with the functions

$$\begin{aligned} f_n(t) &= \sum_{k=0}^{\infty} \binom{n+k}{k} \frac{t^k}{(1+t)^{n+k+1}} f\left(\frac{k}{k+1}\right), \\ f_n(t) &= e^{-nt} \sum_{k=0}^{\infty} \frac{(nt)^k}{k!} f\left(\frac{k}{n}\right), \\ f_n(t) &= \frac{1}{(n-1)!} \int_0^{\infty} f(x) \left(\frac{nx}{t}\right)^{n-1} e^{-nx/t} \frac{n}{t} dx. \end{aligned}$$

Conclude that $f_n(t) \rightarrow f(t)$ uniformly in every finite interval in each of the three cases.
[Hint: Identify $F_{n,t}$ in each case. No calculations are necessary.]

10. *The weak law for stationary sequences.* Suppose $\{X_j, j \in \mathbb{Z}\}$ is a stationary sequence of zero-mean variables. In other words, for any selection of integers N, K , and j_1, \dots, j_K , the distributions of $(X_{j_k}, 1 \leq k \leq K)$ and $(X_{j_k+N}, 1 \leq k \leq K)$ are the same. Define U_j and V_j via the Khinchin truncation (2.2). If $\text{Cov}(U_0, U_n) \rightarrow 0$ as $n \rightarrow \infty$ then $P\{n^{-1}|X_1 + \dots + X_n| > \epsilon\} \rightarrow 0$.

11. *Cantelli's strong law.* Using Theorem 1.1 with $g(x) = x^4$, show directly by using a Chebyshev-style argument that the strong law of large numbers holds if the independent summands come from a common distribution possessing a fourth moment.

12. Show that the series $\sum_{k=1}^{\infty} c_k \sin 2\pi 2^k t$ converges a.e. if $\sum_k c_k^2$ converges. This is a special case of a famous theorem of Kolmogorov. [Hint: Observe that $r_k(t) = \text{sgn } \sin 2\pi 2^{k-1} t$ and attempt to modify the argument for Rademacher series.]

The theory of fair games:

13. *Another fair game.* The random variables X_1, X_2, \dots are independent and assume integral values $k \geq 2$ with probabilities $p_k = c/(k^2 \log k)$ where $c = 1.651 \dots$ is a norming constant. Show that the game with accumulated entrance fees $\alpha_n = cn \log \log n$ is fair in the classical sense.

14. *A fair game where the gambler is guaranteed to lose.* The scheming god Loki, son of Odin, has opened a series of iniquitous gambling dens in the fair city of Odense, Denmark to rival the booming business in St. Petersburg. The game played in these dens of depravity proceeds in a sequence of i.i.d. trials $\{X_j, j \geq 1\}$. In any given trial the gambler wins 2^k Kroner with probability $p_k = [2^k k(k+1)]^{-1}$ for $k \geq 1$ and wins 0 Kroner with probability $p_0 = 1 - \sum_{k \geq 1} p_k$. Show that the expected gain is unit,

$E(X_j) = 1$. To attract the wary mathematician, Loki charges a fair accumulated entrance fee $a_n = n$ to play the game n times, and soon the city of Odense is booming with the guttural sounds of mathematicians in the throes of gambling fever. Suppose a gambler plays through n trials and accumulates winnings $S_n = \sum_{j=1}^n X_j$. While the game is now ostensibly “fair”, show that $P\{S_n - n < -\frac{(1-\epsilon)n}{\log_2 n}\} \rightarrow 1$ as $n \rightarrow \infty$ so that, in the long run, the gambler is guaranteed to lose money. The moral? Don’t gamble with gods—they cheat. This example is due to W. Feller.²² [Hint: Set $a_n = n/\log_2 n$ and introduce the truncated variables $U_j = X_j 1_{(-\infty, a_n]}(X_j)$. (a) Estimate $P\{U_1 = X_1, \dots, U_n = X_n\}$. (b) Show that, for every $\epsilon > 0$, we have $\frac{1}{\log_2 n} < 1 - E(U_1) \leq \frac{1+\epsilon}{\log_2 n}$ for large enough n . (c) By induction or otherwise show that $\sum_{k=1}^r \frac{2^k}{k^2} \leq 9 \frac{2^r}{r^2}$ for each $r \geq 1$. Hence find a good upper bound for $\text{Var}(U_1)$. (d) Show that $P\{|U_1 + \dots + U_n - n E(U_1)| < \frac{\epsilon n}{\log_2 n}\} \rightarrow 1$ by Chebyshev’s inequality.]

15. *The horse race.* A bookie offers 3-for-1 odds on each horse in a three-horse race. (This means that if a gambler bets \$1 on a given horse he will get \$3 if the horse wins and \$0 if it doesn’t.) The race is run repeatedly with the probabilities that the individual horses win unchanging from race to race; the winning horse $X \in \{1, 2, 3\}$ in a given race has distribution $(p_1, p_2, p_3) = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$. The gambler begins with wealth $W_0 = 1$ and invests it all on bets on the three horses in the proportion $f = (f(1), f(2), f(3))$ where $f(1) + f(2) + f(3) = 1$. After the race he reinvests all his winnings in bets in the same proportion f between the horses on the next race and continues in this fashion. His wealth after n races have been run is then given by $W_n = 3W_{n-1}f(X_n)$. Show that $\frac{1}{n} \log W_n$ converges in probability to a limiting value $W(f)$ specified by the investment strategy (or portfolio) f only and find the optimal investment strategy f^* . How does the gambler’s wealth grow (or diminish) under this optimal strategy?

16. *Continuation.* If, instead, the gambler puts all his money on the most likely winner, horse 1, on each race, show that he will go broke with probability one.

17. *The St. Petersburg game with shares.*²³ In the St. Petersburg game the payoff X is equal to 2^k with probability 2^{-k} for $k = 1, 2, \dots$. Suppose a gambler can buy a share of this game each time it is played. If he invests 1 unit in the game, he receives $1/c$ of a share and a return X/c ; for an investment of α , he receives α/c of a share and a return $\alpha X/c$; and, with an investment of c , he receives a full share and a return X . Thus, c represents the price of a share in this game. Suppose X_1, X_2, \dots are independent payoffs in repeated trials in the St. Petersburg game. The gambler begins with one unit of wealth which he invests in the game (and gets a return X_1/c). He then reinvests all his wealth to buy shares in the next game, and proceeds by reinvesting all his wealth each time. Let W_n denote his wealth after n games have been played. Here $W_0 = 1$ and $W_n = W_{n-1}X_n/c$ for $n \geq 1$. Show that there exists a value c^* such that, with probability one, $W_n \rightarrow \infty$ if $c < c^*$ and $W_n \rightarrow 0$ if $c > c^*$. Determine the “fair share price” c^* . [Hint: $W_n = 2^n(\frac{1}{n} \log_2 W_n)$.]

²²W. Feller, “Note on the law of large numbers and ‘fair’ games”, *Annals of Mathematical Statistics*, vol. 16, pp. 301–304, 1945.

²³R. Bell and T. M. Cover, “Competitive optimality of logarithmic investment”, *Mathematics of Operations Research*, vol. 5, pp. 161–166, 1980.

18. Continuation, the cautious gambler. A wary gambler keeps a fixed proportion $\bar{f} = 1 - f$ of his money at each turn in the St. Petersburg game and reinvests the rest. His wealth after the n th trial then satisfies $W_n = W_{n-1}(1 - f + fX_n/c)$. Show that $\frac{1}{n} \log_2 W_n \rightarrow^P W(f, c) = \sum_{k=1}^{\infty} 2^{-k} \log_2 \left(1 - f + \frac{f2^k}{c}\right)$.

19. Continuation. Let $W^*(c) = \max_{0 \leq f \leq 1} W(f, c)$. Determine the value of the entry fee c for which the optimising value f^* drops below 1.

Vapnik–Chervonenkis theory, uniform convergence:

20. Open sets. Show that the family of open sets in the unit interval has infinite Vapnik–Chervonenkis dimension.

21. Closed discs. The family of all closed discs in the Euclidean plane shatters any three-point sample (x_1, x_2, x_3) which is not collinear. But the discs can pick out no more than 15 out of the 16 possible subsamples of a four-point sample (x_1, x_2, x_3, x_4) .

22. Closed rectangles. The family of axis-parallel rectangles in the Euclidean plane can shatter some configurations of four points but no configuration of five points.

23. Convex sets. We say that a set in the Euclidean plane is *convex* if it contains all line segments joining any two points in the set. Show that the family of closed convex sets in the plane has infinite Vapnik–Chervonenkis dimension. [Hint: Consider arrangements on a circle.]

24. Quadrants and orthants. The family \mathfrak{Q} of quadrants in the plane picks out at most $1 + n/2 + n^2/2$ subsamples from any sample of n points. Find a configuration for which this bound is achieved. Conclude that $V_{\mathfrak{Q}} = 2$. What if we deal with the family \mathfrak{Q} of orthants in three (or, more generally, v) dimensions?

25. Half-spaces, Schläfli's theorem.²⁴ Identify $\mathcal{X} = \mathbb{R}^v$ and \mathfrak{H} with the family of half-spaces $\mathbb{H}_w = \{x : xw^\top \geq 0\}$ as w ranges over all points in \mathbb{R}^v . Show that the growth function $D_{\mathfrak{H}}(n) = g(v, n)$ is determined solely by v and n . Demonstrate the validity of the recurrence $g(v, n) = g(v-1, n) + g(v-1, n-1)$ and, hence, by the obvious boundary conditions, $g(1, n) = g(v, 1) = 2$, deduce that $g(v, n) = 2 \sum_{k=0}^{v-1} \binom{n-1}{k}$. Conclude that \mathfrak{H} is a Vapnik–Chervonenkis class and hence that the family of half-spaces is learnable. [Hint: The following beautiful geometrical argument is due to J. G. Wendel.²⁵ With points in general position (every subset of v points is linearly independent) the hyperplanes orthogonal to x_1, \dots, x_{n-1} partition \mathbb{R}^v into $g(v, n-1)$ components; the half-spaces engendered by vectors w in any component pick out the same subsample of (x_1, \dots, x_{n-1}) . These components are of two types: those of type (i), say I in number, do not intersect the hyperplane orthogonal to x_n , and those of type (ii), say J in number, intersect the hyperplane orthogonal to x_n . Type (ii) components engender two new components when x_n is added to the mix; the number of components engendered by type (i) components is unchanged. Thus, $g(v, n) = I + 2J = g(v, n-1) + J$. Each type (ii) component engenders a unique component in the hyperplane [($v-1$)-dimensional subspace] orthogonal to x_n and so $J = g(v-1, n-1)$.]

²⁴L. Schläfli, *Gesammelte Mathematische Abhandlungen I*. Basel, Switzerland: Verlag Birkhäuser, pp. 209–212, 1950.

²⁵J. G. Wendel, “A problem in geometric probability”, *Mathematica Scandinavica*, vol. 11, pp. 109–111, 1962.

26. Continuation, affine half-spaces, hyperplanes. The natural extension of rays in one dimension to \mathbb{R}^v is the family $\mathfrak{H}_{\text{affine}}$ of affine half-spaces defined by sets of the form $\mathbb{H}_{w,t} = \{x : xw^\top \geq t\}$ where the *weight vector* w varies over \mathbb{R}^v and t is a real *threshold*. The equation $xw^\top = t$ determines a hyperplane in v dimensions, each hyperplane inducing two affine half-spaces, one on either side of it. Determine a recurrence for the growth function of this family. What are the boundary conditions? Solve the recurrence and thence determine $V_{\mathfrak{H}_{\text{affine}}}$.

27. Continuation, quadratic forms. The family \mathfrak{Q} of subsets of the plane of the form $ax^2 + bxy + cy^2 + dx + ey + f \geq 0$ includes all closed discs, ellipsoids, and half-spaces. Show that its Vapnik–Chervonenkis dimension is 6. [Hint: The map $(x, y) \mapsto (x^2, xy, y^2, x, y, 1)$ takes \mathbb{R}^2 into \mathbb{R}^6 . The half-spaces in \mathbb{R}^6 induce the subsets on the plane comprising the family \mathfrak{Q} . This is a special case of the general theorem contained in the next problem.]

28. Function spaces, Steele's theorem.²⁶ Suppose \mathcal{F} is a finite-dimensional vector space of real functions on the space \mathcal{X} . Let \mathfrak{G} be the family of sets of the form $\{x : g(x) \geq 0\}$ for $g \in \mathcal{F}$. If \mathcal{F} has dimension v then $V_{\mathfrak{G}} \leq v$. [Hint: Choose any set $\{x_1, \dots, x_v, x_{v+1}\}$ of $v+1$ distinct points from \mathcal{X} . The map $L: f \mapsto (f(x_1), \dots, f(x_v), f(x_{v+1}))$ is linear, whence LF is a linear subspace of \mathbb{R}^{v+1} of dimension at most v . There then exists in \mathbb{R}^{v+1} a non-zero vector w which is orthogonal to LF . In particular, $\sum_j w_j g(x_j) = 0$ for each $g \in \mathcal{F}$. We may suppose that at least one $w_j < 0$ (else consider $-w$ instead). Attempting to pick out precisely those x_j for which $w_j \geq 0$ gives a contradiction.]

29. Improving the rate of convergence in the Glivenko–Cantelli theorem. Symmetrising over the *entire* double sample $(\mathbf{X}, \mathbf{X}')$ in the second symmetrisation step instead of symmetrising by pairwise exchanges improves the rate of convergence. Suppose Π now denotes a random permutation of the double sample $(\mathbf{X}, \mathbf{X}')$. Write $F_n(t; \Pi)$ and $F'_n(t; \Pi)$ for the empirical d.f.s of the first and second samples in the permuted sequence. Let $\rho_n = \sup_t |F_n(t) - F'_n(t)|$ and $\rho_n(\Pi) = \sup_t |F_n(t; \Pi) - F'_n(t; \Pi)|$. By leveraging Problem 3, improve the rate of convergence in the Glivenko–Cantelli theorem to

$$\mathbf{P}\{\sup_t |F_n(t) - F(t)| \geq \epsilon\} \leq 2 \mathbf{P}\{|\rho_n| \geq \epsilon/2\} = \mathbf{P}\{|\rho_n(\Pi)| \geq \epsilon/2\} \leq 4(2n+1)e^{-\epsilon^2 n/8}.$$

30. Continuation, sample size in the Vapnik–Chervonenkis theorem. Conclude that the rate of convergence in (12.1) can be improved to

$$\mathbf{P}\{\sup_{A \in \mathfrak{A}} |F_n(A) - F(A)| \geq \epsilon\} \leq 4D_{\mathfrak{A}}(2n)e^{-\epsilon^2 n/8}. \quad (13.1)$$

If the Vapnik–Chervonenkis class \mathfrak{A} has finite dimension $V_{\mathfrak{A}} = v$, conclude in view of Theorem 12.2 that the upper bound is no larger than δ if $n \geq \frac{16}{\epsilon^2} (v \log \frac{16v}{\epsilon^2} - \log \frac{\delta}{4})$.

31. Continuation, further improvements. The double sample is not technically necessary and the rates can be improved further with some tweaking. Pick a tiny $\alpha > 0$ and let $\mathbf{X}' = (X'_1, X'_2, \dots, X'_{\alpha n})$ be a second sample of size αn which is a small fraction of the size of the original sample. No matter, $F'_n(t)$ should again be close to $F_n(t)$, eventually. First improve the hypergeometric tail bound of Problem 3 to apply to samples of unequal size and then retrace the steps in the proof of the Glivenko–Cantelli theorem to argue that the exponent on the right in (13.1) can be essentially doubled.

²⁶J. M. Steele, *Combinatorial Entropy and Uniform Limit Laws*. Ph.D. thesis, Stanford University, 1975.

XVII

From Inequalities to Concentration

Inequalities form a pillar of the theory of probability. In many settings exact computations are either not required or, more usually, simply not available. An inequality that can be derived in such a situation makes a virtue out of necessity and, in any case, a crude answer is preferable to none. Inequalities, however, can do much more than provide approximate answers. A crisp two-sided inequality can lay the foundation for a precise limit theorem by showing that the quantities of interest are actually concentrated in value near one point. The venerable laws of large numbers arise from such considerations.

C 1, 3–5
A 2, 7–10
§ 6

In the last few decades of the twentieth century V. D. Milman realised that the phenomenon of concentration exhibited in the laws of large numbers was actually much more pervasive than had hitherto been realised and actively promoted investigation of the idea of concentration of measure. In a startling denouement in 1995, M. Talagrand produced a sweeping set of results of almost unbelievable delicacy which showed that concentration was at work in problems across a huge spectrum. In this chapter we shall see how, by strengthening the basic inequalities of the previous chapter, the laws of large numbers can be remarkably extended.

1 Exponential inequalities

Its wide scope and utility notwithstanding, the fondest adherent of Chebyshев's inequality would not claim that it was sharp. Much more precise bounds can be obtained by replacing the bounding quadratic in Chebyshev's inequality by an exponential.

The exponential function $g(x) = e^{\lambda x}$ dominates the Heaviside function $H_0(x)$ for every $\lambda \geq 0$; whence $H_0(x - t) \leq g(x - t) = e^{\lambda(x-t)}$ for all t . Suppose X is any random variable drawn from some distribution F . As a simple variant on the theme of Theorem XVI.1.1 we hence obtain $P\{X \geq t\} \leq E e^{\lambda(X-t)}$ for every choice of $\lambda \geq 0$ and all t .

THEOREM 1 *The tail bound $\mathbf{P}\{X \geq t\} \leq \inf_{\lambda \geq 0} \mathbf{E}(e^{\lambda(X-t)})$ holds for all t .*

Write $M(\lambda) = \mathbf{E}(e^{\lambda X}) = \int_{\mathbb{R}} e^{\lambda x} dF(x)$. Then $M(\lambda)$ is an extended real-valued function of the real variable λ which is finite at least for $\lambda = 0$. We may interpret M as a two-sided Laplace transform of the distribution and, in view of the discussion in Section XV.1, we identify M with the *moment generating function* of the distribution. In this terminology, the moment generating function provides a tail estimate for the distribution via $\mathbf{P}\{X \geq t\} \leq \inf_{\lambda \geq 0} e^{-\lambda t} M(\lambda)$.

EXAMPLE 1) *Another tail bound for the normal.* If X has a standard normal distribution its moment generating function is given by

$$M(\lambda) = \mathbf{E}(e^{\lambda X}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x - x^2/2} dx = e^{\lambda^2/2}$$

by completing squares in the exponent and integrating out the normal density that results. The minimum of the function $f(\lambda) = e^{-\lambda t + \lambda^2/2}$ occurs at $\lambda = t$ and, for $t \geq 0$, Theorem 1 hence gives $\mathbf{P}\{X \geq t\} = 1 - \Phi(t) \leq e^{-t^2/2}$ for the right tail of the normal. That the rate of decay is of the right order may be seen by comparing the bound with those of Lemma VI.1.3 and Theorem X.1.1. While the latter two results are slightly stronger it is notable how effortlessly Theorem 1 produces the right exponential order. ►

As $M(0) = 1$, at worst Theorem 1 provides the trivial bound of 1 for the tail probability. Improved bounds beckon if the moment generating function is finite in some interval about the origin. The subtlety of the theorem rests in the choice of the variable parameter λ which allows one to select the exponential to optimally “fit the distribution”. This is most apparent when we consider sums.

Suppose X_1, X_2, \dots are independent random variables drawn from a common distribution F . Write $M(\lambda) = \mathbf{E}(e^{\lambda X_j}) = \int_{\mathbb{R}} e^{\lambda x} dF(x)$ for the moment generating function of F ; for any given $\lambda \geq 0$, we again interpret $M(\lambda)$ as $+\infty$ if the integral diverges. As usual, we now consider the partial sums $S_n = X_1 + \dots + X_n$ with distribution F^{*n} .

CHERNOFF'S INEQUALITY *For any t we have the tail inequalities*

$$\mathbf{P}\{S_n \geq t\} \leq \left(\inf_{\lambda \geq 0} e^{-\lambda t/n} M(\lambda) \right)^n \text{ and } \mathbf{P}\{S_n \leq -t\} \leq \left(\inf_{\lambda \geq 0} e^{-\lambda t/n} M(-\lambda) \right)^n.$$

PROOF: Fix any $\lambda \geq 0$. The variables $e^{\lambda X_1}, \dots, e^{\lambda X_n}$ are independent, hence uncorrelated, whence

$$\mathbf{E}(e^{\lambda S_n}) = \mathbf{E}\left(\exp \sum_{j=1}^n \lambda X_j\right) = \mathbf{E}\left(\prod_{j=1}^n \exp(\lambda X_j)\right) = \prod_{j=1}^n \mathbf{E} \exp(\lambda X_j) = M(\lambda)^n.$$

Applying Theorem 1 to the variable S_n we hence obtain

$$\mathbf{P}\{S_n \geq t\} \leq \inf_{\lambda \geq 0} e^{-\lambda t} M(\lambda)^n = \left(\inf_{\lambda \geq 0} e^{-\lambda t/n} M(\lambda) \right)^n,$$

it being permissible to take the infimum inside the power as the function $(\cdot)^n$ is monotone. We may obtain a similar result for the left tail by the simple expedient of replacing X_j by $-X_j$ and observing that the moment generating function of the negated variable is $M(-\lambda)$. ▶

Herman Chernoff exhibited his exponentially decreasing bound for the tails of the partial sums S_n in 1952.¹ Following publication it was immediately clear that his exponential bound in t provided a much greater control of the tails of sums than does Chebyshev's inequality which only provides for an inverse quadratic decay in t . The binomial provides a splendid test case.

EXAMPLE 2) *An exponential tail bound for the binomial.* Suppose X_1, X_2, \dots is a sequence of Bernoulli trials with success probability p , the corresponding moment generating function being $M(\lambda) = E(e^{\lambda X_j}) = q + pe^\lambda$. (As usual in this context, $q = 1 - p$.) A straightforward differentiation shows again that the function $f(\lambda) = e^{-\lambda t/n} M(\lambda)$ achieves its minimum value at $\lambda = \log\left(\frac{qt/n}{p(1-t/n)}\right)$ which is positive for $p \leq t/n \leq 1$. Consequently, for $pn \leq t \leq n$,

$$\min_{\lambda \geq 0} f(\lambda) = \left(\frac{t/n}{p} \right)^{-t/n} \left(\frac{1-t/n}{q} \right)^{-(1-t/n)}$$

and by taking logarithms of both sides we may massage the expression into the form $\min_{\lambda} f(\lambda) = e^{-D(t/n, p)}$ where, for $0 \leq r, s \leq 1$, in a slight abuse of notation,

$$D(r, s) = D(Bernoulli(r), Bernoulli(s)) = r \log \frac{r}{s} + (1-r) \log \frac{1-r}{1-s}$$

denotes the Kullback–Leibler divergence between the Bernoulli distributions with success probabilities r and s , respectively [see (XIV.3.2)]. We hence obtain an exponential tail bound for the binomial; the setting crops up frequently enough to make it worth enshrining.

COROLLARY Suppose X_1, \dots, X_n is a sequence of Bernoulli trials with success probability p . Let $S_n = X_1 + \dots + X_n$. If $t \geq pn$, then

$$\mathbf{P}\{S_n \geq t\} = \sum_{k \geq t} \binom{n}{k} p^k (1-p)^{n-k} \leq e^{-nD(t/n, p)}. \quad (1.1)$$

¹H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

As $D(t/n, p)$ is positive by Theorem XIV.3.4, the bound decays exponentially whenever $t > pn$. In the special case when $p = 1/2$ the bound takes the form

$$\mathbf{P}\{S_n \geq t\} = \sum_{k \geq t} \binom{n}{k} 2^{-n} \leq \exp[-n\{\log(2) - h(\frac{t}{n})\}] \quad (1.1')$$

where $h(x) = -x \log(x) - (1-x) \log(1-x)$ is the famous *binary entropy function* of C. E. Shannon.

Asymptotics. If in (1.1') we set $t = n/2 + \xi/2$ to consider the deviations from the mean, we have $\mathbf{P}\{S_n \geq \frac{1}{2}(n+\xi)\} \leq \exp[-n\{\log(2) - h(\frac{1}{2} + \frac{\xi}{2n})\}]$. Expanding the logarithm in a Taylor series and collecting terms we find

$$h(\frac{1}{2} + \frac{\xi}{2n}) = \log(2) - \frac{1}{2} \log(1 - \frac{\xi^2}{n^2}) - \frac{\xi}{2n} \log(\frac{1+\xi/n}{1-\xi/n}) = \log(2) - \frac{\xi^2}{2n} + \mathcal{O}(\frac{\xi^4}{n^3}).$$

Set $\xi = a\sqrt{2n}$ and introduce the standardised zero-mean, unit-variance variable $S_n^* = (S_n - n/2)/\sqrt{n/2}$. Then $\mathbf{P}\{S_n^* \geq a\} \leq e^{-a^2/2 + \mathcal{O}(a^4/n)}$ and the large deviation theorem of Section VI.7 shows that the exponent is of the right order when $a = a_n = o(n^{1/4})$. We hence have more confirmation that Chernoff's bound is exponentially tight at least in the neighbourhood of the mean. ▶

While Chernoff's bound provides a nice framework for setting up exponential bounds, the moment generating function is not easy to work with analytically. Additional structure in the distribution of the random summands can be exploited to reduce the scope of the minimisation to analytically malleable forms at the cost of weakening the exponent in the bound. The hope, of course, is that the simplicity of the resulting function makes up for the loss in precision. The simplest of these settings occurs in the context of the binomial. If she hasn't done so already, the reader should now take the opportunity to read the material in small print in Section XVI.1. If she now compares (1.1') above with (XVI.1.1) she will get a preview in an elementary but important setting of the simplification in the bound that is available at the cost of a little precision. A versatile family of such bounds, of which (XVI.1.1) is a special case, was discovered by Wassily Hoeffding when the random variables are bounded.²

HOEFFDING'S INEQUALITY Suppose $\{X_j, j \geq 1\}$ is a sequence of independent, zero-mean, bounded random variables with $-a_j \leq X_j \leq b_j$ for each j . Let $S_n = X_1 + \dots + X_n$ be the partial sums. Then the left and right tails of the partial sums may be bounded by a common exponential envelope and, for every $t \geq 0$, we have

$$\max\{\mathbf{P}\{S_n \geq t\}, \mathbf{P}\{S_n \leq -t\}\} \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^n (b_j + a_j)^2}\right).$$

²W. Hoeffding, "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963.

PROOF: Suppose $-a < 0 < b$ and let $x = p(-a) + (1-p)b$ be a convex combination of $-a$ and b . Then $p = (b-x)/(b+a)$ and as the exponential function is convex it follows by Jensen's inequality that

$$e^{\lambda x} = e^{\lambda[p(-a)+(1-p)b]} \leq pe^{-\lambda a} + (1-p)e^{\lambda b} = \left(\frac{b-x}{b+a}\right)e^{-\lambda a} + \left(\frac{x+a}{b+a}\right)e^{\lambda b}.$$

Suppose X is a zero-mean, bounded random variable with $-a \leq X \leq b$. As expectation is monotone and linear and $E(X) = 0$, it follows that $E(e^{\lambda X}) \leq f(\lambda)$ where $f(\lambda) = (be^{-\lambda a} + ae^{\lambda b})/(b+a)$. Repeated differentiation shows additionally that $f'(\lambda) = ab(-e^{-\lambda a} + e^{\lambda b})/(b+a)$ and $f''(\lambda) = ab(ae^{-\lambda a} + be^{\lambda b})/(b+a)$. We observe that f and f'' are strictly positive, $f(0) = 1$ and $f'(0) = 0$. Based on the evidence of the examples of this section we should try to show that $f(\lambda) \leq e^{c\lambda^2}$ for some $c \geq 0$. This suggests focusing on the exponent and hence the function $g(\lambda) = \log f(\lambda)$. Differentiating via the chain rule shows that $g'(\lambda) = f'(\lambda)/f(\lambda)$ and $g''(\lambda) = [f''(\lambda)f(\lambda) - f'(\lambda)^2]/f(\lambda)^2$. As $g(0) = g'(0) = 0$, the Taylor expansion of g through two terms shows that, for some λ^* between 0 and λ , we have

$$g(\lambda) = g(0) + g'(0)\lambda + g''(\lambda^*)\lambda^2/2 = g''(\lambda^*)\lambda^2/2,$$

and thus $f(\lambda) = \exp(g''(\lambda^*)\lambda^2/2)$. The exponent will be of the right order if we can show that g'' is bounded. Indeed, after substitution of the expressions for the derivatives of f , straightforward, if slightly tedious, algebraic manoeuvres allow us to wrestle the expression for g'' into the form

$$g''(\lambda) = \frac{(b+a)^2}{2} \cdot \frac{2abe^{\lambda(b-a)}}{(be^{-\lambda a} + ae^{\lambda b})^2}.$$

The second fraction on the right looks complex but is just a concealed expression of the form $2xy/(x+y)^2$, and as $(x+y)^2 \geq 4xy$ for any choice of x and y , it follows that $g''(\lambda) \leq (b+a)^2/4$ for all λ . We hence obtain the simple bound $E(e^{\lambda X}) \leq e^{(b+a)^2\lambda^2/8}$. The rest of the proof follows the same pattern as for Chernoff's bound. By exploiting the independence of the X_j we obtain

$$\begin{aligned} P\{S_n \geq t\} &\leq \inf_{\lambda \geq 0} e^{-\lambda t} E \exp\left(\sum_{j=1}^n \lambda X_j\right) = \inf_{\lambda \geq 0} e^{-\lambda t} \prod_{j=1}^n E(e^{\lambda X_j}) \\ &\leq \inf_{\lambda \geq 0} \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{j=1}^n (b_j + a_j)^2\right) = \exp\left(\frac{-2t^2}{\sum_{j=1}^n (b_j + a_j)^2}\right), \end{aligned}$$

the final step following by a simple minimisation by differentiation. If we replace X_j by $-X_j$ then we see that the left tail $P\{S_n \leq -t\}$ has the same bound as that for the right tail, concluding the proof. ►

In the case of symmetrically bounded variables, $a_j = b_j = \alpha$ for each j , Hoeffding's inequality yields $P\{|S_n| \geq t\} \leq 2e^{-t^2/(2n\alpha^2)}$. Viewed in the proper scale $t = \xi\alpha$ for deviations from the origin, we see that we recover the deviation bound (XVI.1.1') for a simple random walk, reaffirming once more the

SLOGAN *The main features of general random walks are already on display in, and illumined by, the simple Bernoulli random walk.*

The reader has already seen this imbedding make an appearance previously in the construction of Brownian motion (Section X.12) and in the analysis of fluctuations (Section XV.8). We see shades here of Pólya's wise aphorism that we should begin an attack on a proposed problem with the simplest version that we cannot solve.

We resort to exponential bounds when precision in tail estimates is important. Examples appear in the following sections.

2 Unreliable transcription, reliable replication

A collection \mathcal{C} of m binary strings $X_1, \dots, X_m \in \{0, 1\}^n$, called a *code book*, represents encoded information that is to be replicated. Given a binary string X to be replicated, an unreliable transcription mechanism $X \mapsto Y$ produces a noisy copy Y with each bit inverted with probability $p \leq 1/2$ and replicated correctly with probability $q = 1 - p$. We may model the process by introducing a sequence of Bernoulli trials $Z = (Z_1, \dots, Z_n)$, each $Z_k \sim \text{Bernoulli}(p)$, where success connotes error, that is to say, an inverted bit. Then $Y = X + Z$ where addition is to be taken to be modulo 2 componentwise: $Y_k = (X_k + Z_k) \bmod 2$. Given the code book, it is natural to attempt to correct errors in a given transcribed sequence Y (with the originating sequence X obscured) via the maximum likelihood map $Y \mapsto \hat{X}$ which ascribes to Y that sequence \hat{X} in the code book which is closest to Y . The reader has seen the maximum likelihood principle make an appearance in similar contexts in Sections II.10, VIII.3, and XIV.9.

The natural metric in this setting is the *Hamming distance* $\rho(x, y) = \text{card}\{k : x_k \neq y_k\}$ which gives the number of components in which two sequences x and y differ as a measure of the distance between them. The maximum likelihood estimate is hence given by $\hat{X} = X_{j^*}$ where $j^* = \arg \min_j \rho(X_j, Y)$ and ties are broken in any convenient fashion, say lexicographic.

We are interested in the probability that, given a randomly selected information string X in the code book to be replicated, the replication process $X \mapsto Y \mapsto \hat{X}$ results in an error. While this depends strongly on the particular choice of code book, symmetrising the problem by randomising over the choice of the code book yields insights.

Suppose X_1, \dots, X_m are selected by independent sampling from the uniform distribution on the vertices $\{0, 1\}^n$ of the unit cube in n dimensions.

Let \mathbf{X} be a randomly selected element of the code book. By conditioning on the selected element, we have

$$\mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X}\} = \frac{1}{m} \sum_{i=1}^m \mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X} | \mathbf{X} = \mathbf{X}_i\} = \mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X} | \mathbf{X} = \mathbf{X}_1\} \quad (2.1)$$

as the distribution of the code book is invariant with respect to permutation of its elements. Now, conditioned on $\mathbf{X} = \mathbf{X}_1$, the maximum likelihood estimate is in error if any of the elements $\mathbf{X}_2, \dots, \mathbf{X}_m$ is closer to \mathbf{Y} than \mathbf{X}_1 . Accordingly,

$$\mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X} | \mathbf{X} = \mathbf{X}_1\} \leq \sum_{j \geq 2} \mathbf{P}\{\hat{\mathbf{X}} = \mathbf{X}_j | \mathbf{X} = \mathbf{X}_1\} = (m-1) \mathbf{P}\{\hat{\mathbf{X}} = \mathbf{X}_2 | \mathbf{X} = \mathbf{X}_1\}$$

by again exploiting the symmetry of the sampling distribution. Let $K = \{k : X_{1k} \neq X_{2k}\}$ be the set of indices on which X_1 and X_2 disagree. Then the event $\rho(\mathbf{Y}, \mathbf{X}_2) \leq \rho(\mathbf{Y}, \mathbf{X}_1)$ occurs if, and only if, there are errors in transcription in at least $\text{card}(K)/2$ of the bit locations in the set K . Conditioning on the random set K simplifies the computational task. Suppose, say, that $K = \{1, \dots, v\}$. If $\mathbf{X} = \mathbf{X}_1$, the event $\rho(\mathbf{Y}, \mathbf{X}_2) \leq \rho(\mathbf{Y}, \mathbf{X}_1)$ occurs if, and only if, $\sum_{k=1}^v Z_k \geq v/2$. This event has binomial tail probability $\sum_{k \geq v/2} \binom{v}{k} p^k q^{v-k}$ which, in view of Chernoff's bound (1.1) for the binomial, is bounded above by $e^{-vD(1/2, p)}$ where

$$D(1/2, p) = \frac{1}{2} \log \frac{1/2}{p} + \frac{1}{2} \log \frac{1/2}{q} = -\log \sqrt{4pq}.$$

The symmetry inherent in the situation shows that the bound is unaltered for any set K of cardinality v . It follows that

$$\begin{aligned} \mathbf{P}\{\hat{\mathbf{X}} = \mathbf{X}_2 | \mathbf{X} = \mathbf{X}_1\} &= \sum_{v=0}^n 2^{-n} \binom{n}{v} \mathbf{P}\{\hat{\mathbf{X}} = \mathbf{X}_2 | \mathbf{X} = \mathbf{X}_1, \text{card}(K) = v\} \\ &\leq 2^{-n} \sum_{v=0}^n \binom{n}{v} e^{-vD(1/2, p)} = 2^{-n} (1 + e^{-D(1/2, p)})^n = 2^{-n(1 - \log_2(1 + \sqrt{4pq}))}. \end{aligned}$$

The utility of Chernoff's inequality is clear in the exponential decay of the bound; Chebyshev's inequality only provides a trivial bound of order $1/n$. Pooling bounds, we obtain

$$\mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X}\} \leq m \mathbf{P}\{\hat{\mathbf{X}} = \mathbf{X}_2 | \mathbf{X} = \mathbf{X}_1\} \leq 2^{-n(1 - \log_2(1 + \sqrt{4pq}) - \frac{1}{n} \log_2 m)}.$$

The quantity $r = \frac{1}{n} \log_2 m$ represents the amount of information replicated *per bit*. If $r < 1 - \log_2(1 + \sqrt{4pq})$ then the average error probability in the transcription and recovery process $\mathbf{X} \mapsto \mathbf{Y} \mapsto \hat{\mathbf{X}}$ goes exponentially to zero and a strictly positive amount of information per bit can be replicated *reliably*. In particular, for any $\epsilon > 0$, we have $\mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X}\} < \epsilon$ for sufficiently large n . By (2.1) this means that the conditional probabilities $\mathbf{P}\{\hat{\mathbf{X}} \neq \mathbf{X} | \mathbf{X} = \mathbf{X}_i\}$ cannot

exceed 2ϵ for at least half the sequences in the code book. Throwing away the remaining sequences yields a new code book of at least $m/2$ elements each of which can be replicated faithfully with error probability no larger than 2ϵ . As $\frac{1}{n} \log_2 \frac{m}{2} = \frac{1}{n} \log_2 m - \frac{1}{n}$, the information rate r is barely affected.

To summarise, suppose $\delta > 0$ is given, $1 - \delta$ representing a confidence parameter. Say that a vertex $x \in \{-1, 1\}^n$ can be *reliably replicated* if, with $X = x$, the replication process $X \mapsto Y \mapsto \hat{X}$ reproduces x with probability at least $1 - \delta$. We say that a code book C of binary strings can be *reliably replicated* if every string X in it can be reliably replicated. Write $r_* = 1 - \log_2(1 + \sqrt{4pq})$.

THEOREM *If $r < r_*$ then, for all sufficiently large n , there exists a code book of $\lfloor 2^{nr} \rfloor$ elements that can be reliably replicated.*

A more refined analysis can improve the rate r [see Problems 10–14] but our analysis already points to a remarkable result that we shall be satisfied with.

SLOGAN *It is possible to reliably replicate a strictly positive amount of information per bit through an unreliable transcription process.*

This is the discrete analogue of Shannon's capacity theorem for the Gaussian channel discussed in Section XIV.9.

3 Concentration, the Gromov–Milman formulation

We are grown to a blasé acceptance of the law of large numbers by frequent exposure and this perhaps has inured us to its remarkable nature. In its simplest form it may be expressed (roughly) in the following unexceptionable statement:

In a long run of coin tosses it is likely that roughly half show heads.

If the reader wishes she can add probabilistic flesh to these verbal bones:

If S_n denotes the partial sums of a sequence of symmetric Bernoulli trials then $P\{|S_n - n/2| \geq t\} \leq 2e^{-2t^2/n}$ for all $t \geq 0$.

This is, of course, Hoeffding's inequality (XVI.1.1). The more precise view says that S_n has likely deviations only of order \sqrt{n} around its mean of $n/2$ or, with the proper scaling, that $\frac{1}{n}S_n$ is concentrated at $1/2$.

In a more general context, suppose X_1, X_2, \dots is a sequence of independent random variables drawn from a common distribution F . If F has finite expectation, we may express the law of large numbers, roughly speaking, in the following statement, which, as in the case of coin tosses, has a suitable probabilistic interpretation:

The function $f(X_1, \dots, X_n) = \frac{1}{n}(X_1 + \dots + X_n)$ is essentially constant.

This is a remarkable assertion. A function of many, independently varying coordinates acts for all practical purposes as if it were concentrated at one point! Could this be part of a more general phenomenon?

In investigations beginning in 1971, V. D. Milman and M. Gromov realised that the concentration phenomenon that is captured by the law of large numbers is much more widespread than had originally been realised. The next two and a half decades saw a growing appreciation of the ubiquity of concentration of measure. This was capped in 1995 by a *tour de force* of M. Talagrand which not only provided the proper framework for viewing a growing patchwork of partial results but vastly extended them. The results may be summarised in a pithy slogan of almost unbelievable scope.

SLOGAN *An honest function which depends smoothly on many independent variables is essentially constant.*

In many ways the slogan provides a quite remarkable closure with some of the earliest ideas in probability. And, as we shall see in the applications, it reduces problems which had originally required enormous ingenuity to a routine application of the concentration idea. In the following sections we shall produce refined and precise versions of the slogan buttressed by beautiful applications. In this we could do no better than follow Talagrand’s groundbreaking work.³

We begin with an abstract space \mathcal{X} as the basic coordinate space. We equip \mathcal{X} with a probability measure μ (on a suitable σ -algebra of sets). We shall deal with the product space $\Omega = \mathcal{X}^n$ equipped with product measure $\mathbf{P} = \mu^{\otimes n}$.

Now it would be absurd to assert that a general function f on Ω is constant without further clarification. If the concentration phenomenon is to be in evidence at all then surely the function f must be honest, depending on *all* the coordinate variables and not excessively on any subgroup, and must have a certain degree of smoothness to eschew wild behaviour. A natural way to capture smoothness is through the *Lipschitz property* by equipping Ω with a metric or distance ρ .⁴ Thus, we interpret $\rho(x, y)$ as the distance between points x and y in Ω . If \mathbb{A} is any subset of Ω we also write $\rho(x, \mathbb{A})$ to mean the infimum of the values $\rho(x, y)$ as y ranges over all points in \mathbb{A} .

In order to make measurability questions transparent it will suffice to assume that Ω is separable and complete, that is to say, Ω contains a countably dense subset of points and every Cauchy sequence converges. Such spaces are called *Polish spaces*; \mathbb{R}^n provides a typical example.

³M. Talagrand, “Concentration of measure and isoperimetric inequalities in product spaces”, *Publications Math. IHES*, vol. 81, pp. 73–205, 1995.

⁴The reader will recall that a metric or distance ρ in the space Ω is a real-valued map on $\Omega \times \Omega$, $(x, y) \mapsto \rho(x, y)$, satisfying the following properties: (1) symmetry, $\rho(x, y) = \rho(y, x)$, (2) positivity, $\rho(x, y) \geq 0$ with equality if, and only if, $x = y$, and (3) the triangle inequality, $\rho(x, y) \leq \rho(x, z) + \rho(y, z)$.

DEFINITION A function f on Ω is *Lipschitz* if there is a constant c (the *Lipschitz constant*) such that $|f(x) - f(y)| \leq c\rho(x, y)$ for all $x, y \in \Omega$.

If f is Lipschitz it is patently continuous (indeed, uniformly continuous) and hence Borel measurable. We may simplify considerations a little by the observation that, if f has Lipschitz constant c , then f/c has Lipschitz constant 1. We may hence just as well deal with the class of Lipschitz functions with constant 1. Suppose accordingly that f is Lipschitz (with Lipschitz constant 1). If $f(X_1, \dots, X_n)$ is to be essentially constant what would a reasonable guess for the constant be? Some measure of central tendency is surely suggested. As f may not have expectation we consider as candidate for the constant any *median* M_f , that is to say, any real value M_f such that $P\{f \leq M_f\} \geq 1/2$ and $P\{f \geq M_f\} \geq 1/2$.⁵

While functions are more natural objects, rephrasing the setting in terms of sets leads to a cleaner abstract formulation. Let $\mathbb{A} = \{f \leq M_f\}$ be the set of points $x \in \Omega$ for which $f(x) \leq M_f$. For every $t > 0$, we define the *t-fattening* of \mathbb{A} as the set $\mathbb{A}_t = \{x : \rho(x, \mathbb{A}) \leq t\} = \{x : \inf\{\rho(x, y) : y \in \mathbb{A}\} \leq t\}$. Figure 1 should not be taken too literally but should help provide some visual intuition. Now suppose $x \in \mathbb{A}_t$. Then, by the Lipschitz property,

$$f(x) \leq \inf\{f(y) + \rho(x, y) : y \in \mathbb{A}\} \leq M_f + \inf\{\rho(x, y) : y \in \mathbb{A}\} \leq M_f + t.$$

It follows that $P\{f \leq M_f + t\} \geq P(\mathbb{A}_t)$ or, equivalently, $P\{f > M_f + t\} \leq 1 - P(\mathbb{A}_t)$.

By a similar line of argument, if we set $\mathbb{B} = \{f \geq M_f\} = \{x : f(x) \geq M_f\}$ and let $\mathbb{B}_t = \{x : \rho(x, \mathbb{B}) \leq t\}$ be its t-fattening, then, if $x \in \mathbb{B}_t$, we have

$$f(x) \geq \sup\{f(y) - \rho(x, y) : y \in \mathbb{B}\} \geq M_f - \inf\{\rho(x, y) : y \in \mathbb{B}\} \geq M_f - t.$$

It follows that $P\{f \geq M_f - t\} \geq P(\mathbb{B}_t)$ or, equivalently, $P\{f < M_f - t\} \leq 1 - P(\mathbb{B}_t)$.

The sets \mathbb{A} and \mathbb{B} are “high-probability” sets each of probability at least one-half. The probability that f deviates from a median value may now be bounded by the probabilities outside the t-fattening of these high-probability sets by

$$P\{|f - M_f| > t\} \leq 2 \max\{1 - P(\mathbb{A}_t), 1 - P(\mathbb{B}_t)\}.$$

While it is difficult to infer much about the structure of a given high-probability set, the situation becomes more tractable if we symmetrise by considering all of them. For $t \geq 0$, we define the *concentration function* $\alpha(t) = \alpha(t; \rho, P)$ as the least upper bound of the values $1 - P(\mathbb{C}_t)$ as \mathbb{C} ranges over all high-probability sets of probability at least one-half.

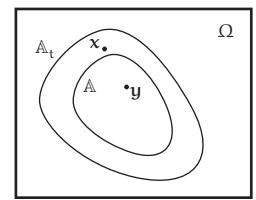


Figure 1: t-fattening.

⁵In applications there is usually not much difference between the mean and the median but the median is more convenient and natural here. See Problems XIV.14,15 and Problem 1 of this chapter.

THEOREM Suppose f has Lipschitz constant 1. Then $\mathbf{P}\{|f - M_f| > t\} \leq 2\alpha(t)$ for every $t \geq 0$.

If f has Lipschitz constant c then the bound becomes $\mathbf{P}\{|f - M_f| > t\} \leq 2\alpha(t/c)$ by a simple scale of function.

To calculate the probability that f deviates from a median value it suffices hence to estimate the concentration function, that is to say, the probability outside the immediate vicinity of high-probability sets. Our next step is to figure out a reasonable candidate for a metric.

4 Talagrand views a distance

We begin by introducing a nonce notation for vector norms in \mathbb{R}^n . To complement the usual Euclidean 2-norm $\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_n^2}$ we introduce the 1-norm $|\mathbf{x}| = |x_1| + \dots + |x_n|$, the number of vertical bars providing a visual prompt as to which norm is meant. If $\mathbf{x} \in \{0, 1\}^n$ then the 1-norm of \mathbf{x} simplifies to $|\mathbf{x}| = \sum_{j=1}^n x_j$ and we may identify $|\mathbf{x}|$ as the number of 1s in the sequence x_1, \dots, x_n . As before, vector inequalities in \mathbb{R}^n are to be interpreted component-wise.

What is a good choice of metric in the Gromov–Milman formulation? While it is hard to imagine a natural metric in an abstract space, product spaces come equipped with a natural notion of distance: the *Hamming distance* $\rho(\mathbf{x}, \mathbf{y}) = \text{card}\{j : x_j \neq y_j\}$ between points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ is the number of coordinates on which \mathbf{x} differs from \mathbf{y} .

The Hamming distance is not a very potent notion of distance but serves as a natural starting point for investigation. If we write $\mathbf{h}(\mathbf{x}, \mathbf{y}) \in \{0, 1\}^n$ as the Hamming vector whose components $h(x_j, y_j) = 1\{x_j \neq y_j\}$ are the indicators for whether the corresponding components of \mathbf{x} and \mathbf{y} agree or not, we may write the Hamming metric in the form $\rho(\mathbf{x}, \mathbf{y}) = |\mathbf{h}(\mathbf{x}, \mathbf{y})| = \sum_j h(x_j, y_j)$, each mismatched component contributing once to the sum. This additive representation suggests an extension by implementing a selective weighting of the contribution of mismatched elements. Suppose $\mathbf{r} \geq \mathbf{0}$ is a vector of positive components. We define the *r-Hamming distance* by

$$\rho_{\mathbf{r}}(\mathbf{x}, \mathbf{y}) = \sum_{j: x_j \neq y_j} r_j = \sum_j r_j h(x_j, y_j) = \langle \mathbf{r}, \mathbf{h}(\mathbf{x}, \mathbf{y}) \rangle, \quad (4.1)$$

where we may identify the sum in the penultimate expression with the natural inner product $\langle \mathbf{r}, \mathbf{h}(\mathbf{x}, \mathbf{y}) \rangle = \mathbf{r}\mathbf{h}(\mathbf{x}, \mathbf{y})^\top$ in \mathbb{R}^n . In this formulation we identify ordinary Hamming distance $\rho(\mathbf{x}, \mathbf{y})$ with $\rho_1(\mathbf{x}, \mathbf{y}) = \sum_j h(x_j, y_j)$. It is easy enough to verify that $\rho_{\mathbf{r}}$ is a bona fide metric on Ω though it is much less clear what a good selection of \mathbf{r} would be other than to set it to be 1. When at an impasse one can do worse than turn to the tosses of a coin which, as Talagrand reaffirms, forms an inexhaustible source of inspiration.

FROM ONE-POINT TO MANY-POINT APPROXIMATION

Suppose $\mathbf{X} = (X_1, \dots, X_n)$ is a sequence of Bernoulli trials with success probability 1/2. Suppose, for definiteness, that n is even. Let $\mathbb{A} = \{\mathbf{y} : |\mathbf{y}| \leq n/2\}$ be the collection of points $\mathbf{y} \in \{0, 1\}^n$ with no more than $n/2$ 1s. Then, for every $t > 0$, we have

$$\mathbf{P}\{\rho_1(\mathbf{X}, \mathbb{A}) > t\} = 1 - \mathbf{P}(\mathbb{A}_t) = \mathbf{P}\{|\mathbf{X}| > \frac{n}{2} + t\} \leq e^{-2t^2/n} \quad (4.2)$$

by Hoeffding's inequality (XVI.1.1). We may view (4.2) as a statement of the degree of approximation of the realisation \mathbf{x} of a random point by one point in the set \mathbb{A} , any point \mathbf{y} in \mathbb{A} that is "closest" to \mathbf{x} . If $|\mathbf{x}| \leq n/2$ then approximation is trivial as \mathbf{x} itself is in \mathbb{A} . If $|\mathbf{x}| > n/2$ on the other hand the best we can do is select as approximation for \mathbf{x} a point $\mathbf{y} \leq \mathbf{x}$ exactly $n/2$ of whose components are 1. Which point? Suppose $|\mathbf{x}| = n/2 + t$. Then there are precisely $\binom{n/2+t}{n/2}$ points \mathbf{y} , each containing exactly $n/2$ 1s, which are closest to \mathbf{x} in \mathbb{A} . These points \mathbf{y} constitute the best approximations for \mathbf{x} in \mathbb{A} . Each of these approximations "miss" \mathbf{x} in exactly t coordinates but they don't all "miss" the same coordinates. The reader may now be willing to entertain the idea that the average of these approximations may do a better job of approximating \mathbf{x} than any one alone.

Following this speculative line of thought, in a slight abuse of notation consider the *mean Hamming error vector* $\mathbf{h}(\mathbf{x}) = \mathbf{h}(\mathbf{x}, \mathbb{A})$ obtained by averaging $\mathbf{h}(\mathbf{x}, \mathbf{y})$ componentwise over all $\mathbf{y} \leq \mathbf{x}$ with $|\mathbf{y}| = n/2$. Writing $h_j = h(\mathbf{x})_j$ for the components of $\mathbf{h}(\mathbf{x})$, it is clear that $h_j = 0$ if $x_j = 0$ while, if $x_j = 1$,

$$h_j = \binom{n/2+t-1}{n/2} / \binom{n/2+t}{n/2} = \frac{t}{n/2+t}.$$

Thus, if $|\mathbf{x}| > n/2$ then $\mathbf{h}(\mathbf{x}) = \left(\frac{|\mathbf{x}|-n/2}{|\mathbf{x}|}\right)\mathbf{x}$. As, for each $\mathbf{x} \in \{0, 1\}^n$, $\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_n^2} = \sqrt{x_1 + \dots + x_n} = \sqrt{|\mathbf{x}|}$, it follows that the (Euclidean) 2-norm of the mean error in approximation may be bounded by

$$\|\mathbf{h}(\mathbf{x})\| = \left(\frac{|\mathbf{x}|-n/2}{|\mathbf{x}|}\right) \|\mathbf{x}\| = \frac{|\mathbf{x}|-n/2}{\sqrt{|\mathbf{x}|}} \leq \frac{|\mathbf{x}|-n/2}{\sqrt{n/2}}$$

whenever $|\mathbf{x}| > n/2$. On the other hand, $\mathbf{h}(\mathbf{x}) = 0$, trivially, if $|\mathbf{x}| \leq n/2$ so that the expression on the right bounds $\|\mathbf{h}(\mathbf{x})\|$ for all $\mathbf{x} \in \{0, 1\}^n$. We now have the tantalising possibility of measuring the "distance" from \mathbb{A} by the length of the mean error in approximation by (many) points in \mathbb{A} . Writing $\rho_0(\mathbf{x}, \mathbb{A}) = \|\mathbf{h}(\mathbf{x})\|$ to make the connection with distance explicit, we have

$$\rho_0(\mathbf{x}, \mathbb{A}) \leq \frac{(|\mathbf{x}|-n/2)^+}{\sqrt{n/2}}$$

for all $x \in \{0, 1\}^n$. If X represents a point selected randomly from the vertices $\{0, 1\}^n$, we hence obtain by Hoeffding's inequality (XVI.1.1) that

$$\mathbf{P}\{\rho_0(X, A) > t\} \leq \mathbf{P}\left\{|X| \geq \frac{n}{2} + t\sqrt{\frac{n}{2}}\right\} \leq e^{-t^2}. \quad (4.2')$$

Compared with (4.2) we see that the factor n^{-1} has disappeared in the exponent! This suggests that $\rho_0(X, A) = \|h(X)\|$ is *very* concentrated around 0, or, in an alternative way of saying the same thing, that the entire space is essentially concentrated within a small neighbourhood of the set A . Approximation by many points has reduced the error very significantly vis à vis one-point approximation. While we could certainly build more ammunition for this viewpoint by, say, looking next at Bernoulli trials with success probability p , and then generalising the setting further, the possibility of high concentration by many-point approximation is already seductive enough for us to plunge in.

TALAGRAND'S CONVEX DISTANCE

The space $\Omega = \mathcal{X}^n$ is not necessarily a metric space. Building on coin-tossing intuition, we plan to reduce considerations to Euclidean space \mathbb{R}^n to allow of geometric characterisations. For any subset A of Ω and any x in Ω , let $U'_A(x)$ denote the collection of distinct binary strings $h(x, y)$ as y ranges over all the points of A . Thus, $U'_A(x)$ consists of all the distinct Hamming error vectors obtained in approximating x by points y in A . In order to consider averages of these error vectors it will be helpful to “fatten” $U'_A(x)$ to include generic averages. We recall that a linear combination of the form $z = \alpha_1 h_1 + \cdots + \alpha_m h_m$ of a finite set of points h_1, \dots, h_m in \mathbb{R}^n is a *convex combination* if $\alpha_1, \dots, \alpha_m$ are positive and add to one.

DEFINITION Let x be any point and A any subset of $\Omega = \mathcal{X}^n$. Then *Talagrand's convex distance* between x and A , denoted $\rho_0(x, A)$, is defined to be the smallest value of $\|z\|$ as z ranges over all convex combinations of points in $U'_A(x)$.

The set comprised of all convex combinations of points in $U'_A(x)$ is called the *convex hull* of $U'_A(x)$ and denoted $\text{ch } U'_A(x)$. Then, in a compact notation, we may write $\rho_0(x, A) = \min\{\|z\| : z \in \text{ch } U'_A(x)\}$.

The alert reader may wonder why the minimum is achievable. She should argue that it is by: (1) observing that $U'_A(x)$ is a subset of $\{0, 1\}^n$, hence patently a finite set; (2) arguing that the convex hull of a finite set of points in \mathbb{R}^n is a closed and bounded set; and (3) confirming that the 2-norm $z \mapsto \|z\|$ is a continuous function whence it achieves its minimum on $\text{ch } U'_A(x)$ by (the version for \mathbb{R}^n of) Theorem XXI.2.2.

In order to get a better feel for the nature of the convex distance, we will look at it from two different perspectives.

The set of successors. While it is most natural to deal with the set of Hamming error vectors $U'_A(x)$ and its convex hull $\text{ch } U'_A(x)$, a variation on the

theme is analytically more convenient. Suppose \mathbf{g}, \mathbf{h} are vertices in $\{0, 1\}^n$. We say that \mathbf{g} is a *successor* of \mathbf{h} if $\mathbf{h} \leq \mathbf{g}$. We now enlarge $U'_A(x)$ to form the set $U_A(x)$ of all successors of the points of $U'_A(x)$. It is clear that $U'_A(x) \subseteq U_A(x)$ (as any element \mathbf{h} of $U'_A(x)$ is trivially its own successor) so that $\text{ch } U'_A(x) \subseteq \text{ch } U_A(x)$. It follows that $\min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U'_A(x)\} \geq \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U_A(x)\}$. We will now see that equality actually obtains.

It is intuitive that the convex hulls of $U'_A(x)$ and $U_A(x)$ inherit the successor relationship. Indeed, suppose $\mathbf{z} = \sum_k \alpha_k \mathbf{g}_k$ is a convex combination of points in $U_A(x)$. For each k , there exists $\mathbf{y}_k \in A$ such that \mathbf{g}_k is a successor of $\mathbf{h}_k = \mathbf{h}(x, \mathbf{y}_k)$ in $U'_A(x)$, that is to say, $\mathbf{h}(x, \mathbf{y}_k) \leq \mathbf{g}_k$. It follows that $\mathbf{z} = \sum_k \alpha_k \mathbf{g}_k \geq \sum_k \alpha_k \mathbf{h}(x, \mathbf{y}_k) = \mathbf{z}'$ as each α_k is positive. As \mathbf{z}' is a convex combination of vectors in $U'_A(x)$, it follows that each vector in $\text{ch } U_A(x)$ componentwise dominates some vector in $\text{ch } U'_A(x)$. And thus, $\min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U'_A(x)\} \leq \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U_A(x)\}$. As we've established the inequality in both directions, we must have

$$\rho_0(x, A) = \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U'_A(x)\} = \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U_A(x)\} \quad (4.3)$$

and enlarging the set $U'_A(x)$ to form the set of successors $U_A(x)$ does not alter the distance of the corresponding convex hulls from the origin.

The connection with r-Hamming distance. In view of the identification in (4.1) of the r -Hamming distance with the inner product in Euclidean space \mathbb{R}^n , the r -Hamming distance of x from A is given by $\rho_r(x, A) = \min\{\langle \mathbf{r}, \mathbf{y} \rangle : \mathbf{y} \in A\} = \min\{\langle \mathbf{r}, \mathbf{h} \rangle : \mathbf{h} \in U'_A(x)\}$. [We may replace the infimum by a minimum as the inner products $\langle \mathbf{r}, \mathbf{h} \rangle$ range over only a finite number of possibilities as \mathbf{h} ranges through $U'_A(x)$.] We now argue that the minimum value of the inner product on the right is unaffected if we allow \mathbf{h} to range over the larger set $\text{ch } U'_A(x)$.

The inner product $\bar{r}: \mathbf{h} \mapsto \langle \mathbf{r}, \mathbf{h} \rangle$ is a continuous linear function on \mathbb{R}^n equipped with the Euclidean 2-norm. As $\text{ch } U'_A(x)$ is a closed and bounded set, \bar{r} achieves its minimum value in $\text{ch } U'_A(x)$, say, at a point $\hat{\mathbf{z}}$. Then $\hat{\mathbf{z}} = \sum_k \alpha_k \mathbf{h}_k$ is a convex combination of points \mathbf{h}_k in $U'_A(x)$. Consequently,

$$\min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U'_A(x)\} = \langle \mathbf{r}, \hat{\mathbf{z}} \rangle = \langle \mathbf{r}, \sum_k \alpha_k \mathbf{h}_k \rangle = \sum_k \alpha_k \langle \mathbf{r}, \mathbf{h}_k \rangle$$

by linearity of inner product. There must hence exist at least one k with $\langle \mathbf{r}, \mathbf{h}_k \rangle \leq \langle \mathbf{r}, \hat{\mathbf{z}} \rangle$ from which it follows that $\min\{\langle \mathbf{r}, \mathbf{h} \rangle : \mathbf{h} \in U'_A(x)\} \leq \min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U'_A(x)\}$. But the expression on the right minimises over a larger set and so is certainly no larger than the expression on the left. It follows that the two expressions are indeed equal to each other and the minimum is unaffected by moving to the convex hull: $\min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in U'_A(x)\} = \min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U'_A(x)\}$.

It will be convenient to expand the set of minimisation a little further. As each element \mathbf{g} of $U_A(x)$ is a successor of, hence componentwise dominates, some element \mathbf{h} of $U'_A(x)$, it follows that $\langle \mathbf{r}, \mathbf{g} \rangle \geq \langle \mathbf{r}, \mathbf{h} \rangle$ and so we may replace $U'_A(x)$ by $U_A(x)$ without changing the minimum value of the inner product. And now by the same argument as in the previous paragraph, we can

further expand considerations from the set $U_{\mathbb{A}}(x)$ to its convex hull $\text{ch } U_{\mathbb{A}}(x)$ again without changing the minimum value of the inner product. And thus $\min\{\langle \mathbf{r}, \mathbf{g} \rangle : \mathbf{g} \in U_{\mathbb{A}}(x)\} = \min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)\}$. We hence obtain

$$\rho_r(x, \mathbb{A}) = \min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U'_{\mathbb{A}}(x)\} = \min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)\}.$$

All is in readiness to relate the r -Hamming distance to the convex distance. The Cauchy–Schwarz inequality tells us that

$$\min\{\langle \mathbf{r}, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U'_{\mathbb{A}}(x)\} \leq \|\mathbf{r}\| \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U'_{\mathbb{A}}(x)\} = \|\mathbf{r}\| \rho_0(x, \mathbb{A}).$$

Normalising \mathbf{r} to unit length, it follows that $\rho_r(x, \mathbb{A}) \leq \rho_0(x, \mathbb{A})$ for every choice of $\mathbf{r} \geq 0$ with $\|\mathbf{r}\| = 1$ and, consequently,

$$\max\{\rho_r(x, \mathbb{A}) : \mathbf{r} \geq 0, \|\mathbf{r}\| = 1\} \leq \rho_0(x, \mathbb{A}). \quad (4.4)$$

The reader may find it satisfying that the convex distance dominates all (weight-normalised) Hamming distances. But we can squeeze a little more water from this stone.

Suppose \mathbf{z}^* is a minimum-length vector in the convex hull of successors $\text{ch } U_{\mathbb{A}}(x)$: $\|\mathbf{z}^*\| = \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)\} = \rho_0(x, \mathbb{A})$. Then, for any $\mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)$ and any $0 \leq \lambda \leq 1$, the convex combination $\lambda\mathbf{z} + (1-\lambda)\mathbf{z}^* = \mathbf{z}^* + \lambda(\mathbf{z} - \mathbf{z}^*)$ is in $\text{ch } U_{\mathbb{A}}(x)$. As \mathbf{z}^* has minimum length, we have

$$\|\mathbf{z}^*\|^2 \leq \|\mathbf{z}^* + \lambda(\mathbf{z} - \mathbf{z}^*)\|^2 = \|\mathbf{z}^*\|^2 + 2\lambda\langle \mathbf{z}^*, \mathbf{z} - \mathbf{z}^* \rangle + \lambda^2\|\mathbf{z} - \mathbf{z}^*\|^2$$

by expanding out the square of the norm using linearity of inner product. It follows that $2\lambda\langle \mathbf{z}^*, \mathbf{z} - \mathbf{z}^* \rangle + \lambda^2\|\mathbf{z} - \mathbf{z}^*\|^2 \geq 0$ and as this inequality must hold for all choices of positive λ , we must have $0 \leq \langle \mathbf{z}^*, \mathbf{z} - \mathbf{z}^* \rangle = \langle \mathbf{z}^*, \mathbf{z} \rangle - \|\mathbf{z}^*\|^2$ for all $\mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)$. The case when the convex hull of $U_{\mathbb{A}}(x)$ includes the origin is trivial; accordingly suppose that $\text{ch } U_{\mathbb{A}}(x)$ does not include the origin. We may now normalise by setting $\mathbf{r}^* = \mathbf{z}^*/\|\mathbf{z}^*\|$ to form a unit-length positive vector. We then obtain $\langle \mathbf{z}, \mathbf{z}^* \rangle = \|\mathbf{z}^*\|\langle \mathbf{z}, \mathbf{r}^* \rangle \geq \|\mathbf{z}^*\|^2$ for all $\mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)$. It follows that $\rho_{\mathbf{r}^*}(x, \mathbb{A}) = \min\{\langle \mathbf{r}^*, \mathbf{z} \rangle : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)\} \geq \|\mathbf{z}^*\| = \rho_0(x, \mathbb{A})$, and, *a fortiori*,

$$\max\{\rho_r(x, \mathbb{A}) : \mathbf{r} \geq 0, \|\mathbf{r}\| = 1\} \geq \rho_0(x, \mathbb{A}). \quad (4.4')$$

In view of (4.4), equality actually obtains.

Pooling our findings (4.3, 4.4, 4.4') yields several ways of looking at the convex distance.

THEOREM 1 *Talagrand's convex distance may be written in any of the following equivalent forms:*

$$\begin{aligned} \rho_0(x, \mathbb{A}) &= \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U'_{\mathbb{A}}(x)\} \\ &= \min\{\|\mathbf{z}\| : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(x)\} = \max\{\rho_r(x, \mathbb{A}) : \mathbf{r} \geq 0, \|\mathbf{r}\| = 1\}. \end{aligned}$$

In a sense the convex distance “optimises” over the choices of weights in the r -Hamming distance. Another way of looking at it is that this choice of distance “fattens” the set \mathbb{A} maximally. This is the reason behind the pronounced concentration in evidence in (4.2').

The reader may now have perhaps begun to feel hopeful about the programme outlined. And indeed very strong concentration obtains.

TALAGRAND'S THEOREM *Suppose \mathbf{X} is drawn from the product measure $\mathbf{P} = \mu^{\otimes n}$ on $\Omega = \mathcal{X}^n$ and \mathbb{A} is a subset of Ω of strictly positive probability. Then, for every $t > 0$, we have*

$$\mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A}) \geq t\} \leq \frac{1}{\mathbf{P}(\mathbb{A})} \exp\left(\frac{-t^2}{4}\right), \quad (4.5)$$

the bound independent of dimension n .

If $\mathbf{P}(\mathbb{A}) \geq 1/2$ then $\mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A}) \geq t\} = 1 - \mathbf{P}(\mathbb{A}_t) \leq 2e^{-t^2/4}$ where \mathbb{A}_t is the t -fattening of \mathbb{A} with respect to the convex distance ρ_0 .

COROLLARY *The Gromov–Milman concentration function with respect to convex distance is bounded by $\alpha(t; \rho_0, \mathbf{P}) \leq 2e^{-t^2/4}$ independent of dimension.*

The Talagrand bound on the concentration function not only doesn't depend on n but dies so quickly with t that we are left with a remarkable consequence.

SLOGAN *Every set of non-negligible probability (say, probability one-half) in a product space has essentially the entire space concentrated in its immediate vicinity.*

This rephrasing of the slogan of the previous section articulates a deep, non-intuitive property of independence and product spaces. The reader may well find the result hard to visualise and even harder to believe. And she may find it unsettling that such prosaic objects as sets of probability one-half have the power to surprise. Here as elsewhere asking the right question is half the battle. Concentration inequalities of this type show that there is a kind of “surface hardening” near the boundary of non-negligible sets. While we shall not pursue the analogy further, the reader may be struck by the apparently serendipitous connection with sphere-hardening and the familiar isoperimetric inequality. This is no accident; the link is provided by the central limit theorem.⁶

The concentration with respect to convex distance presented here is only one of a family of isoperimetric inequalities that arise out of Talagrand's induction method. My defence for limiting the presentation of Talagrand's ideas to the setting of convex distance is that from a pedagogical point of view a novice reader is better served by presenting a topic cohesively in a particular

⁶M. Ledoux, “Les inégalités isopérimétriques en analyse et probabilités”, Séminaire Bourbaki, Astérisque, vol. 216, pp. 343–375, June 1993.

context—a too early introduction of abstraction without context only muddies the waters and hampers a proper appreciation of the basic ideas. Keeping this admonition in mind, the convex distance serves admirably as a vehicle to promote the idea of concentration of measure; it is subtle enough to permit an exhibition of the power of the method and supple enough to fit a variety of deep applications. The reader who masters this material and hungers for more will want to visit Talagrand’s *magnum opus* for a view of the full power of the induction method and the variety of isoperimetric inequalities that it engenders.

Following Talagrand’s proof of his theorem, several alternative proofs have been discovered using varied ideas drawn from martingales, information theory, and logarithmic Sobolev inequalities. But for sheer charm it is hard to match Talagrand’s original proof by induction. And, at least in the opinion of this author, the induction method exposes the machinery so clearly and in such an elementary way that the reader following along can not only know that the proof is correct but can *believe* it.

5 The power of induction

In this section I have tried to emphasise the clarity of the ideas in Talagrand’s theorem by laying out the proof not perhaps in the most logically parsimonious way but as perhaps it might be discovered.

It will be convenient to clear the square-root in the Euclidean 2-norm by taking squares. In order to set the stage for a Chernoff-style bound, we consider

$$\begin{aligned} \mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A}) \geq t\} &= \mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A})^2 \geq t^2\} \\ &\leq \mathbf{E} \exp \lambda (\rho_0(\mathbf{X}, \mathbb{A})^2 - t^2) = e^{-\lambda t^2} \mathbf{E} \exp \lambda \rho_0(\mathbf{X}, \mathbb{A})^2 \end{aligned} \quad (5.1)$$

and we need to estimate $\mathbf{E} \exp \lambda \rho_0(\mathbf{X}, \mathbb{A})^2$. We leave $\lambda \geq 0$ unspecified for now and select an appropriate value later when the best choice becomes apparent.

To set up a proof by induction of Talagrand’s theorem we should begin by verifying the result for the base case $n = 1$.

THE INDUCTION BASE

Suppose \mathbb{A} is a subset of \mathcal{X} . If $x \in \mathbb{A}$ then $0 = h(x, x) \in U'_\mathbb{A}(x)$. In this case $ch U'_\mathbb{A}(x)$ contains the origin, whence $\rho_0(x, \mathbb{A}) = 0$. If, on the other hand, $x \notin \mathbb{A}$, then $1 = h(x, y) \in U'_\mathbb{A}(x)$ for any $y \in \mathbb{A}$ so that $U'_\mathbb{A}(x)$ is a singleton set with 1 as its only element. In this case $ch U'_\mathbb{A}(x) = \{1\}$, whence $\rho_0(x, \mathbb{A}) = 1$. Summarising, $\rho_0(x, \mathbb{A}) = 1_{\mathbb{A}}(x)$ is the indicator for the set \mathbb{A} and, accordingly, we have

$$\mathbf{E} \exp \lambda \rho_0(x, \mathbb{A})^2 = \mathbf{P}(\mathbb{A}) + (1 - \mathbf{P}(\mathbb{A}))e^\lambda \quad (\lambda \geq 0).$$

Experience with Hoeffding’s inequality suggests that the right-hand side does not give a good enough uniform bound in $\mathbf{P}(\mathbb{A})$, especially in the critical region

where λ is small. To compensate for low-probability sets we multiply both sides by $\mathbf{P}(\mathbb{A})$ to obtain

$$\mathbf{P}(\mathbb{A}) \mathbf{E} \exp \lambda \rho_0(x, \mathbb{A})^2 = \mathbf{P}(\mathbb{A}) [\mathbf{P}(\mathbb{A}) + (1 - \mathbf{P}(\mathbb{A})) e^\lambda].$$

We are accordingly led to consider the function $g(p) = p[p + (1 - p)e^\lambda]$ for $0 \leq p \leq 1$. Now

$$g'(p) = e^\lambda - 2p(e^\lambda - 1) \geq e^\lambda - 2(e^\lambda - 1) = 2 - e^\lambda$$

so that $g'(p) \geq 0$ if $0 \leq \lambda \leq \log 2$. Fix λ in this interval. It follows that g increases monotonically from its value $g(0) = 0$ to its value $g(1) = 1$ as p increases from 0 to 1 and *a fortiori* $g(p) \leq 1$ for each p in the unit interval. In particular, setting $p = \mathbf{P}(\mathbb{A})$, we obtain

$$\mathbf{E} \exp \lambda \rho_0(X, \mathbb{A})^2 = \mathbf{P}(\mathbb{A})^{-1} g(\mathbf{P}(\mathbb{A})) \leq \mathbf{P}(\mathbb{A})^{-1}$$

for every choice of λ in the interval $0 \leq \lambda \leq \log 2$. This looks promising. We should now try to set up an induction to prove it in general.

THE INDUCTION STEP

Talagrand's powerful idea was to reduce considerations of \mathcal{X}^n to iterated considerations of \mathcal{X} . As induction hypothesis, fix n and suppose that for any (measurable) subset \mathbb{A}' of \mathcal{X}^n of strictly positive probability we have

$$\mathbf{E} \exp \lambda \rho_0(X, \mathbb{A}')^2 \leq \mathbf{P}(\mathbb{A}')^{-1},$$

with λ to be specified.

Now suppose $\mathbb{A} \subseteq \mathcal{X}^n \times \mathcal{X}$ is a measurable subset of strictly positive probability in the product space in $n + 1$ dimensions. To reduce considerations to the space \mathcal{X}^n , we now introduce two ideas of projections (Figure 2). The *projection* of \mathbb{A} into \mathcal{X}^n is the set $\Pi \mathbb{A}$ of points $y \in \mathcal{X}^n$ for which $(y, \omega) \in \mathbb{A}$ for some $\omega \in \mathcal{X}$. For each $\omega \in \mathcal{X}$ we also define the *projection slice* as the set $\Pi_\omega \mathbb{A}$ of points $y \in \mathcal{X}^n$ for which $(y, \omega) \in \mathbb{A}$.

Begin with the projection slice $\Pi_\omega \mathbb{A}$ and any $x \in \mathcal{X}^n$. Suppose $g \in \{0, 1\}^n$ is a successor of $h(x, y)$ for some $y \in \Pi_\omega \mathbb{A}$. Then $(g, 0) \in \{0, 1\}^{n+1}$ is a successor of $h((x, \omega), (y, \omega)) = (h(x, y), 0)$ as is obvious because the last component is the same in (x, ω) and (y, ω) and we are given that $h(x, y) \leq g$ for the remaining components. Accordingly, if $g \in U_{\Pi_\omega \mathbb{A}}(x)$ then $(g, 0) \in U_{\mathbb{A}}(x, \omega)$. Now suppose $\sigma = \sum_k \alpha_k g_k$ is a convex combination of points

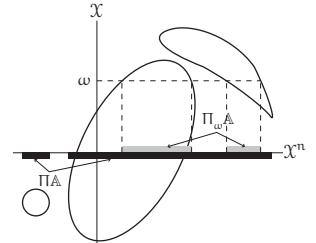


Figure 2: Projections.

$\mathbf{g}_k \in U_{\Pi_\omega \mathbb{A}}(\mathbf{x})$. Then $(\mathbf{g}_k, 0) \in U_{\mathbb{A}}(\mathbf{x}, \omega)$ and it follows that the convex combination $\sum_k \alpha_k (\mathbf{g}_k, 0) = (\sum_k \alpha_k \mathbf{g}_k, 0) = (\boldsymbol{\sigma}, 0)$ is in $\text{ch } U_{\mathbb{A}}(\mathbf{x}, \omega)$. Thus, if $\boldsymbol{\sigma} \in \text{ch } U_{\Pi_\omega \mathbb{A}}(\mathbf{x})$ then $(\boldsymbol{\sigma}, 0) \in \text{ch } U_{\mathbb{A}}(\mathbf{x}, \omega)$.

Consider now the projection $\Pi \mathbb{A}$. Suppose that \mathbf{g} is a successor of $\mathbf{h}(\mathbf{x}, \mathbf{y})$ for some $\mathbf{y} \in \Pi \mathbb{A}$. Then there exists $\omega' \in \mathcal{X}$ such that $(\mathbf{y}, \omega') \in \mathbb{A}$ and $\mathbf{h}((\mathbf{x}, \omega), (\mathbf{y}, \omega')) \leq (\mathbf{g}, 1)$ for all ω as, whatever the values of ω and ω' , the final component can only be a 0 or a 1 and in all cases is dominated by 1.⁷ Accordingly, if $\mathbf{g} \in U_{\Pi \mathbb{A}}(\mathbf{x})$ then $(\mathbf{g}, 1) \in U_{\mathbb{A}}(\mathbf{x}, \omega)$. Now consider any convex combination $\boldsymbol{\tau} = \sum_k \alpha_k \mathbf{g}_k$ of points \mathbf{g}_k in $U_{\Pi \mathbb{A}}(\mathbf{x})$. Then $(\mathbf{g}_k, 1) \in U_{\mathbb{A}}(\mathbf{x}, \omega)$ whence $\sum_k \alpha_k (\mathbf{g}_k, 1) = (\sum_k \alpha_k \mathbf{g}_k, \sum_k \alpha_k) = (\boldsymbol{\tau}, 1)$ is in $\text{ch } U_{\mathbb{A}}(\mathbf{x}, \omega)$. Thus, if $\boldsymbol{\tau} \in \text{ch } U_{\Pi \mathbb{A}}(\mathbf{x})$ then $(\boldsymbol{\tau}, 1) \in \text{ch } U_{\mathbb{A}}(\mathbf{x}, \omega)$ for every $\omega \in \mathcal{X}$.

An induction is now nicely set up reducing considerations of the set \mathbb{A} in $n + 1$ dimensions to considerations of the n -dimensional projections $\Pi \mathbb{A}$ and $\Pi_\omega \mathbb{A}$. Pick any $\boldsymbol{\sigma} \in \text{ch } U_{\Pi_\omega \mathbb{A}}(\mathbf{x})$ and $\boldsymbol{\tau} \in \text{ch } U_{\Pi \mathbb{A}}(\mathbf{x})$. Then $(\boldsymbol{\sigma}, 0)$ and $(\boldsymbol{\tau}, 1)$ are both in the convex hull of $U_{\mathbb{A}}(\mathbf{x}, \omega)$ and hence so is any convex combination

$$\mathbf{z} = (1 - \theta)(\boldsymbol{\sigma}, 0) + \theta(\boldsymbol{\tau}, 1) = ((1 - \theta)\boldsymbol{\sigma} + \theta\boldsymbol{\tau}, \theta)$$

for any $0 \leq \theta \leq 1$. As the square of the Euclidean 2-norm is just the sum of the squares of the coordinates, we obtain

$$\|\mathbf{z}\|^2 = \|(1 - \theta)\boldsymbol{\sigma} + \theta\boldsymbol{\tau}\|^2 + \theta^2. \quad (5.2)$$

To keep the induction going we will want to bound the right-hand side by an expression involving $\|\boldsymbol{\sigma}\|^2$ and $\|\boldsymbol{\tau}\|^2$. The convex combination of $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ that appears on the right holds the key.

LEMMA 1 *The map $\psi: \mathbf{s} \mapsto \|\mathbf{s}\|^2$ which associates with each point $\mathbf{s} \in \mathbb{R}^n$ the square of its Euclidean length is convex.*

PROOF: I will provide a proof by induction in keeping with the theme of this section though the reader who finds the assertion believable without the necessity of a formal proof should skip on to see how it all pans out. (For a direct proof expand out the norm square using linearity of inner product.)

Begin with the base case $n = 1$. Suppose $s, t \in \mathbb{R}$ and $0 \leq \theta \leq 1$. Then

$$((1 - \theta)s^2 + \theta t^2) - ((1 - \theta)s + \theta t)^2 = \theta(1 - \theta)(s - t)^2 \geq 0.$$

As induction hypothesis now suppose that $\|(1 - \theta)\mathbf{s} + \theta\mathbf{t}\|^2 \leq (1 - \theta)\|\mathbf{s}\|^2 + \theta\|\mathbf{t}\|^2$. Consider any pair of points $(\mathbf{s}, s'), (\mathbf{t}, t') \in \mathbb{R}^n \times \mathbb{R}$. Two applications of the induction hypothesis show that

$$\begin{aligned} & \| (1 - \theta)(\mathbf{s}, s') + \theta(\mathbf{t}, t') \|^2 = \| (1 - \theta)\mathbf{s} + \theta\mathbf{t} \|^2 + ((1 - \theta)s' + \theta t')^2 \\ & \leq \{ (1 - \theta)\|\mathbf{s}\|^2 + \theta\|\mathbf{t}\|^2 \} + \{ (1 - \theta)(s')^2 + \theta(t')^2 \} = (1 - \theta)\|(\mathbf{s}, s')\|^2 + \theta\|(\mathbf{t}, t')\|^2 \end{aligned}$$

⁷This is the reason why “bulging” out the set of Hamming error vectors $U'_{\mathbb{A}}(\mathbf{x})$ to form the set of successors of error vectors $U_{\mathbb{A}}(\mathbf{x})$ is useful.

and this completes the induction. ▶

As $\rho_0((x, \omega), \mathbb{A}) = \min\{\|z'\| : z' \in \text{ch } U_{\mathbb{A}}(x, \omega)\}$, by an application of Jensen's inequality to (5.2), it follows that

$$\rho_0((x, \omega), \mathbb{A})^2 \leq \|z\|^2 \leq (1 - \theta)\|\sigma\|^2 + \theta\|\tau\|^2 + \theta^2.$$

As the inequality holds for all choices of $\sigma \in \text{ch } U_{\Pi_\omega \mathbb{A}}(x)$ and $\tau \in \text{ch } U_{\Pi \mathbb{A}}(x)$, we obtain

$$\rho_0((x, \omega), \mathbb{A})^2 \leq (1 - \theta)\rho_0(x, \Pi_\omega \mathbb{A})^2 + \theta\rho_0(x, \Pi \mathbb{A})^2 + \theta^2 \quad (5.3)$$

for all choices of $0 < \theta < 1$. All is in readiness for the induction step.

Suppose $(X, \omega) \in \mathcal{X}^n \times \mathcal{X}$ is picked according to the probability law $P \otimes \mu (= \mu^{\otimes n} \otimes \mu)$. By Fubini's theorem,

$$E \exp \lambda \rho_0((X, \omega), \mathbb{A})^2 = \int_{\mathcal{X}} \left\{ \int_{\mathcal{X}^n} \exp \lambda \rho_0((x, \omega), \mathbb{A})^2 P(dx) \right\} \mu(d\omega).$$

The fundamental induction (5.3) now shows that, for each $\omega \in \mathcal{X}$, the inner integral may be bounded by

$$\begin{aligned} & \int_{\mathcal{X}^n} \exp \lambda \rho_0((x, \omega), \mathbb{A})^2 P(dx) \\ & \leq \int_{\mathcal{X}^n} \exp \lambda ((1 - \theta)\rho_0(x, \Pi_\omega \mathbb{A})^2 + \theta\rho_0(x, \Pi \mathbb{A})^2 + \theta^2) P(dx) \\ & = e^{\lambda \theta^2} \int_{\mathcal{X}^n} (\exp \lambda \rho_0(x, \Pi_\omega \mathbb{A})^2)^{1-\theta} (\exp \lambda \rho_0(x, \Pi \mathbb{A})^2)^\theta P(dx). \end{aligned} \quad (5.4)$$

The integrand certainly looks messy but if the reader steps back and focuses on the flex parameter θ she will recognise that the setting is ripe for an application of Hölder's inequality. If f and g are positive functions and $0 < \theta < 1$, we may write Hölder's inequality in the form $\int fg \leq (\int f^{1/(1-\theta)})^{1-\theta} (\int g^{1/\theta})^\theta$. Identifying $f = \exp(1 - \theta)\lambda \rho_0(x, \Pi_\omega \mathbb{A})^2$ and $g = \exp \theta \lambda \rho_0(x, \Pi \mathbb{A})^2$, the expression on the right-hand side of (5.4) is hence bounded above by

$$\begin{aligned} & e^{\lambda \theta^2} \left\{ \int_{\mathcal{X}^n} \exp \lambda \rho_0(x, \Pi_\omega \mathbb{A})^2 P(dx) \right\}^{1-\theta} \left\{ \int_{\mathcal{X}^n} \exp \lambda \rho_0(x, \Pi \mathbb{A})^2 P(dx) \right\}^\theta \\ & \leq e^{\lambda \theta^2} P(\Pi_\omega \mathbb{A})^{-(1-\theta)} P(\Pi \mathbb{A})^{-\theta} = \frac{1}{P(\Pi \mathbb{A})} \left\{ \left(\frac{P(\Pi_\omega \mathbb{A})}{P(\Pi \mathbb{A})} \right)^{-(1-\theta)} e^{\lambda \theta^2} \right\} \end{aligned} \quad (5.4')$$

by two applications of the induction hypothesis. We are now ready to optimise over the parameters. It will suffice to show that

$$\left(\frac{P(\Pi_\omega \mathbb{A})}{P(\Pi \mathbb{A})} \right)^{-(1-\theta)} e^{\lambda \theta^2} \leq 2 - \frac{P(\Pi_\omega \mathbb{A})}{P(\Pi \mathbb{A})} \quad (5.5)$$

for a suitable choice of λ and any $0 < \theta < 1$. Indeed, if (5.5) holds then

$$\int_{\mathcal{X}^n} \exp \lambda \rho_0((x, \omega), \mathbb{A})^2 P(dx) \leq \frac{1}{P(\Pi \mathbb{A})} \left(2 - \frac{P(\Pi_\omega \mathbb{A})}{P(\Pi \mathbb{A})} \right).$$

Integrating out with respect to ω by Fubini's theorem yields

$$\begin{aligned} E \exp \lambda \rho_0((X, \omega), \mathbb{A})^2 &\leq \int_X \frac{1}{P(\Pi \mathbb{A})} \left(2 - \frac{P(\Pi_\omega \mathbb{A})}{P(\Pi \mathbb{A})} \right) \mu(d\omega) \\ &= \frac{1}{P(\Pi \mathbb{A})} \left(2 - \frac{P \otimes \mu(\mathbb{A})}{P(\Pi \mathbb{A})} \right) = \frac{1}{P \otimes \mu(\mathbb{A})} \cdot \frac{P \otimes \mu(\mathbb{A})}{P(\Pi \mathbb{A})} \left(2 - \frac{P \otimes \mu(\mathbb{A})}{P(\Pi \mathbb{A})} \right). \end{aligned}$$

Identifying $s = P \otimes \mu(\mathbb{A}) / P(\Pi \mathbb{A})$, the elementary inequality $s(2-s) \leq 1$ [which is just a tortuous way of writing $(s-1)^2 \geq 0$] valid for all real s hence shows that $E \exp \lambda \rho_0((X, \omega), \mathbb{A})^2 \leq [P \otimes \mu(\mathbb{A})]^{-1}$, completing the induction.

It only remains to show the validity of (5.5) for an appropriate range of λ . As $P(\Pi_\omega \mathbb{A}) / P(\Pi \mathbb{A}) \leq 1$ we are accordingly led to consider the function $f_\zeta(\theta) = \zeta^{-(1-\theta)} e^{\lambda \theta^2}$ parametrised by the values of ζ in the unit interval $0 \leq \zeta \leq 1$. The appropriate range for λ falls out of the investigation of this function. The proof of the elementary lemma given below is not difficult (though it requires surprising algebraic finesse) but is mildly tedious as it needs a separate examination of cases. It can be skipped on a first reading.

LEMMA 2 *For every $0 \leq \zeta \leq 1$ and $0 \leq \lambda \leq 1/4$, we have $\inf_\theta f_\zeta(\theta) \leq 2 - \zeta$.*

PROOF: Routine calculations show that, for each ζ , the derivatives of $f_\zeta(\theta)$ are given by $f'_\zeta(\theta) = f_\zeta(\theta)(\log \zeta + 2\lambda\theta)$ and $f''_\zeta(\theta) = f_\zeta(\theta)(\log \zeta + 2\lambda\theta)^2 + 2\lambda f_\zeta(\theta) \geq 0$. Bearing in mind that the argument θ must lie in the unit interval, it follows that when $\zeta \geq e^{-2\lambda}$ the function $f_\zeta(\theta)$ has a unique minimum at the interior point $\theta^* = -\log(\zeta)/(2\lambda)$ and when $0 \leq \zeta \leq e^{-2\lambda}$ it achieves its minimum at the boundary point $\theta^* = 1$. We consider these cases in turn.

Suppose $0 \leq \zeta \leq e^{-2\lambda}$. It will suffice to show that $f_\zeta(\theta^*) = e^\lambda \leq 2 - e^{-2\lambda} \leq 2 - \zeta$. Thus it is enough to show that $1 \geq \frac{1}{2}(e^{-2\lambda} + e^\lambda) = e^{-\lambda/2} \cosh(3\lambda/2)$. By comparing corresponding terms of the Taylor series expansions it is easy to see that $\cosh(x) = \frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$. Accordingly, it is enough to show that $1 \geq e^{-\lambda/2} e^{(3\lambda/2)^2/2} = e^{-\frac{1}{2}\lambda + \frac{9}{8}\lambda^2}$. The inequality holds when $\lambda/2 \geq 9\lambda^2/8$ or when $\lambda \leq 4/9$ which is more than we need.

Suppose now that $e^{-2\lambda} \leq \zeta \leq 1$. Then $0 \leq \theta^* = -\log(\zeta)/(2\lambda) \leq 1$. The reader should resist the temptation to attempt to optimise the expression for $f_\zeta(\theta^*)$ directly: it will suffice to show that $f_\zeta(\theta^*) \leq 2 - \zeta$ for λ in a proper range. Writing $e^{\Psi(\zeta)} = (2 - \zeta) / f_\zeta(\theta^*)$ and taking logarithms of both sides to isolate the exponent, we have to show that

$$0 \leq \Psi(\zeta) = \log(2 - \zeta) - \log f_\zeta(\theta^*) = \log(2 - \zeta) + \log \zeta + (\log \zeta)^2 / (4\lambda).$$

As $\psi(1) = 0$ it will suffice to show that $\psi'(\zeta) \leq 0$ for $\zeta \leq 1$ as a decreasing function with a value of 0 at the right endpoint of an interval must be positive on the whole interval. Now $\psi'(\zeta) = -(2-\zeta)^{-1} + \zeta^{-1} + (2\lambda\zeta)^{-1} \log(\zeta)$ so that $\psi'(1) = 0$. It will be convenient to clear the fractions by multiplying throughout by ζ . As $\psi'(\zeta) \leq 0$ if, and only if, $\zeta\psi'(\zeta) \leq 0$ and the value of $\zeta\psi'(\zeta)$ at $\zeta = 1$ is identically 0, it will suffice to show that $(\zeta\psi'(\zeta))' \geq 0$. (A positive second derivative implies an increasing first derivative; if, additionally, the function value at the right endpoint of an interval is zero then it must increase to zero through negative values.) Now $(\zeta\psi'(\zeta))' = -2(2-\zeta)^{-2} + (2\lambda\zeta)^{-1}$ by a routine differentiation. It follows that $(\zeta\psi'(\zeta))' \geq 0$ if, and only if, $4\lambda\zeta \leq 4 - 4\zeta + \zeta^2$, which is the same as saying that $(\zeta - 2(1+\lambda))^2 \geq 4\lambda(2+\lambda)$. Allowing ζ to range over the unit interval, the expression on the left is minimised when $\zeta = 1$. We hence require $(-1 - 2\lambda)^2 \geq 4\lambda(2 + \lambda)$ which means $\lambda \leq 1/4$. ▶

A look back at the induction base shows that we had required $0 \leq \lambda \leq \log 2$ and so we pick the smaller of the intervals for λ and declare victory.

THEOREM Suppose \mathbf{X} is drawn from the product measure $\mathbf{P} = \mu^{\otimes n}$ on $\Omega = \mathcal{X}^n$ and \mathbb{A} is a subset of Ω of strictly positive probability. For every choice of $0 \leq \lambda \leq 1/4$, we have $\mathbf{E} \exp \lambda \rho_0(\mathbf{X}, \mathbb{A})^2 \leq \mathbf{P}(\mathbb{A})^{-1}$ independent of dimension n .

In view of the Chebyshev exponential bound (5.1) this completes the proof of Talagrand's theorem.



6 Sharpening, or the importance of convexity

The reader who steps back and looks at the proof of Talagrand's theorem may feel that the selection of $\lambda = 1/4$ in (5.1) leading to (4.5) has a slightly *ad hoc* flavour to it and wonder if it can be tightened. And indeed it can. Let us retrace our path through the proof and see what can be improved.

The choice $\lambda = 1/4$ in the Chebyshev exponential bound (5.1) led us to work with $e^{\rho_0(\cdot, \mathbb{A})^2/4}$ where $\frac{1}{4}\rho_0(\mathbf{x}, \mathbb{A})^2 = \min_{\mathbf{z} \in \text{ch } U_{\mathbb{A}}(\mathbf{x})} \frac{1}{4}\|\mathbf{z}\|^2$. What is the property of the square of the Euclidean norm that is critical? It is contained in Lemma 1 of the previous section: the convexity of the map $\mathbf{z} \mapsto \frac{1}{4}\|\mathbf{z}\|^2 = \sum_{j=1}^n \frac{1}{4}z_j^2$ is what is key to all that follows. We may then be led to consider what happens if we replace $\frac{1}{4}z_j^2$ by a more flexible convex, increasing function $\xi(z_j)$ and the weighted square of the Euclidean norm $\frac{1}{4}\|\mathbf{z}\|^2 = \sum_{j=1}^n \frac{1}{4}z_j^2$ by a sum of coordinate terms of the form $\Xi(\mathbf{z}) = \sum_{j=1}^n \xi(z_j)$. The square of the convex distance $\frac{1}{4}\rho_0(\mathbf{x}, \mathbb{A})^2$ then gets replaced by $r(\mathbf{x}, \mathbb{A}) = \inf\{\Xi(\mathbf{z}) : \mathbf{z} \in \text{ch } U_{\mathbb{A}}(\mathbf{x})\}$ and we consider the function $e^{r(\mathbf{x}, \mathbb{A})}$ in place of $e^{\rho_0(\mathbf{x}, \mathbb{A})^2/4}$. The objective now is to replace the bound $\mathbf{E} \exp \frac{1}{4}\rho_0(\mathbf{X}, \mathbb{A})^2 \leq \mathbf{P}(\mathbb{A})^{-1}$ by a bound of the form $\mathbf{E} \exp r(\mathbf{X}, \mathbb{A}) \leq \mathbf{P}(\mathbb{A})^{-\alpha}$ for some $\alpha \geq 0$. If the programme is to work as outlined then the function $\xi(z) = \xi_\alpha(z)$ [as also the functions $\Xi(z) = \Xi_\alpha(z)$ and $r(\mathbf{x}, \mathbb{A}) = r_\alpha(\mathbf{x}, \mathbb{A})$] must of course depend upon the choice of α and the question now is what is the best choice of convex function ξ .

To search for a function without any further guidance appears ambitious, to put it mildly, but let us see where the induction leads us. Retracing our path, (5.3) is now

replaced by

$$r((x, \omega), \mathbb{A}) \leq (1 - \theta)r(x, \Pi_\omega \mathbb{A}) + \theta r(x, \Pi \mathbb{A}) + \xi(\theta).$$

We leverage Hölder's inequality as before and (5.4, 5.4') now have as counterparts

$$\begin{aligned} \int_{\mathcal{X}^n} \exp r((x, \omega)) \mathbf{P}(dx) &\leq e^{\xi(\theta)} \int_{\mathcal{X}^n} (\exp r(x, \Pi_\omega \mathbb{A}))^{1-\theta} (\exp r(x, \Pi \mathbb{A}))^\theta \mathbf{P}(dx) \\ &\leq e^{\xi(\theta)} \left\{ \int_{\mathcal{X}^n} \exp r(x, \Pi_\omega \mathbb{A}) \mathbf{P}(dx) \right\}^{1-\theta} \left\{ \int_{\mathcal{X}^n} \exp r(x, \Pi \mathbb{A}) \mathbf{P}(dx) \right\}^\theta \\ &\leq e^{\xi(\theta)} \mathbf{P}(\Pi_\omega \mathbb{A})^{-\alpha(1-\theta)} \mathbf{P}(\Pi \mathbb{A})^{-\alpha\theta} = \frac{1}{\mathbf{P}(\Pi \mathbb{A})^\alpha} \left\{ \left(\frac{\mathbf{P}(\Pi_\omega \mathbb{A})}{\mathbf{P}(\Pi \mathbb{A})} \right)^{-\alpha(1-\theta)} e^{\xi(\theta)} \right\}, \end{aligned} \quad (6.1)$$

the final steps following by induction hypothesis. Reusing notation we are accordingly led to consider the function $f_\zeta(\theta) = \zeta^{-\alpha(1-\theta)} e^{\xi(\theta)}$. As before, the parameter ζ and the convexity variable θ take values in the unit interval. It will suffice if we can show that $\min_\theta f_\zeta(\theta) = 1 + \alpha - \alpha\zeta$ for a suitable choice of function $\xi(\cdot)$. (This contains the bound of Lemma 2 of the previous section when $\alpha = 1$.)

If ξ possesses a second derivative ξ'' then $\xi'' \geq 0$ as ξ is to be convex. We now optimise over θ by taking derivatives of $f_\zeta(\theta)$ and find that the best choice of θ occurs when $\alpha \log \zeta + \xi'(\theta) = 0$ or $\zeta = \exp(-\frac{1}{\alpha} \xi'(\theta))$. As we require $\min_\theta f_\zeta(\theta) = 1 + \alpha - \alpha\zeta$, the function ξ must satisfy

$$\exp((1 - \theta)\xi'(\theta) + \xi(\theta)) = 1 + \alpha - \alpha \exp(-\frac{1}{\alpha} \xi'(\theta)),$$

and taking logarithms to isolate the exponents yields the condition

$$(1 - \theta)\xi'(\theta) + \xi(\theta) = \log(1 + \alpha - \alpha e^{-\xi'(\theta)/\alpha}).$$

Taking derivatives of both sides simplifies the expression dramatically and we obtain

$$(1 - \theta)\xi''(\theta) - \xi'(\theta) + \xi'(\theta) = \frac{\xi''(\theta)e^{-\xi'(\theta)/\alpha}}{1 + \alpha - \alpha e^{-\xi'(\theta)/\alpha}}.$$

Clearing the denominator on the right, solving for $e^{-\xi'(\theta)/\alpha}$, and taking logarithms to isolate the exponent again yields a differential equation as charming as it is unexpected:

$$\xi'(\theta) = \alpha \log(1 + \alpha - \alpha\theta) - \alpha \log(1 - \theta) - \alpha \log(1 + \alpha). \quad (6.2)$$

From here on it only requires routine manipulation. Integrating out we obtain

$$\begin{aligned} \xi(\theta) - \xi(0) &= \alpha \int_0^\theta \log(1 + \alpha - \alpha\nu) d\nu - \alpha \int_0^\theta \log(1 - \nu) d\nu - \alpha \int_0^\theta \log(1 + \alpha) d\nu \\ &= \int_{1+\alpha-\alpha\theta}^{1+\alpha} \log(y) dy - \alpha \int_{1-\theta}^1 \log(y) dy - \alpha\theta \log(1 + \alpha), \end{aligned}$$

the natural changes of variable reducing the integrals on the right to an elementary integral of the logarithm. We may as well set $\xi(0) = 0$ and so, by performing the indicated integrations of the logarithm by parts and collecting terms, we obtain

$$\xi(\theta) = \xi_\alpha(\theta) = \alpha(1 - \theta) \log(1 - \theta) + (1 + \alpha - \alpha\theta) \log\left(\frac{1 + \alpha}{1 + \alpha - \alpha\theta}\right). \quad (6.3)$$

One could have received long odds against the possibility that such an elegant and simple expression would emerge by purely algebraic manoeuvres from such stony ground.

LEMMA 1 *The function $\xi(\theta)$ given by (6.3) is increasing and convex in the unit interval. Moreover, $\xi(\theta) \geq \frac{\alpha}{2(\alpha+1)}\theta^2$ and, a fortiori, $\rho_0(x, \mathbb{A})^2 \leq \frac{2(\alpha+1)}{\alpha}r(x, \mathbb{A})$.*

PROOF: One more differentiation of both sides of (6.2) shows now that

$$\xi''(\theta) = \frac{\alpha}{1-\theta} - \frac{\alpha^2}{1+\alpha-\alpha\theta} = \frac{\alpha}{(1-\theta)(1+\alpha-\alpha\theta)} \geq \frac{\alpha}{1+\alpha} \geq 0.$$

Hence ξ is increasing and convex in the unit interval. Now, by Taylor's theorem, there exists $0 \leq \theta^* \leq \theta$ such that $\xi(\theta) = \xi(0) + \xi'(0)\theta + \frac{1}{2}\xi''(\theta^*)\theta^2 \geq \frac{\alpha}{2(\alpha+1)}\theta^2$ as $\xi(0) = \xi'(0) = 0$. It follows that, for any $z \in \text{ch } U_{\mathbb{A}}(x)$,

$$\|z\|^2 = \sum_{j=1}^n z_j^2 \leq \frac{2(\alpha+1)}{\alpha} \sum_{j=1}^n \xi(z_j) = \frac{2(\alpha+1)}{\alpha} \Xi(z).$$

Minimising over $z \in \text{ch } U_{\mathbb{A}}(x)$ yields $\rho_0(x, \mathbb{A}) \leq \frac{2(\alpha+1)}{\alpha}r(x, \mathbb{A})$. ►

The replacement of $\frac{1}{4}\rho_0(x, \mathbb{A})^2$ by $r(x, \mathbb{A})$ is now completely justified, the analysis magically yielding the function $\xi(\theta) = \xi_{\alpha}(\theta)$ that optimises the key

LEMMA 2 *For every $\alpha \geq 0$ and $0 \leq \zeta \leq 1$, we have $\min_{0 \leq \theta \leq 1} \zeta^{-\alpha(1-\theta)} e^{\xi_{\alpha}(\theta)} = 1 + \alpha - \alpha\zeta$ where $\xi_{\alpha}(\theta)$ is the convex function given by (6.3).*

We may now proceed with the induction as before. We replace (5.5) by

$$\left(\frac{\mathbf{P}(\Pi_{\omega} \mathbb{A})}{\mathbf{P}(\Pi \mathbb{A})} \right)^{-\alpha(1-\theta)} e^{\xi_{\alpha}(\theta)} \leq 1 + \alpha - \alpha \frac{\mathbf{P}(\Pi_{\omega} \mathbb{A})}{\mathbf{P}(\Pi \mathbb{A})}$$

whence integrating out (6.1) with respect to ω yields

$$\mathbf{E} \exp r((X, \omega), \mathbb{A}) \leq \frac{1}{\mathbf{P} \otimes \mu(\mathbb{A})^{\alpha}} \cdot \left(\frac{\mathbf{P} \otimes \mu(\mathbb{A})}{\mathbf{P}(\Pi \mathbb{A})} \right)^{\alpha} \left(1 + \alpha - \alpha \frac{\mathbf{P} \otimes \mu(\mathbb{A})}{\mathbf{P}(\Pi \mathbb{A})} \right).$$

Identifying $\beta = \mathbf{P} \otimes \mu(\mathbb{A}) / \mathbf{P}(\Pi \mathbb{A})$, the induction will be complete once we establish

LEMMA 3 *If $\alpha > 0$ and $\beta > 0$ then $1 + \alpha - \alpha\beta \leq \beta^{-\alpha}$.*

PROOF: The function $x^{-\alpha}$ is convex and its graph lies above its tangent line at $x = 1$ given by the equation $y = 1 + \alpha - \alpha x$. It follows that $x^{-\alpha} \geq 1 + \alpha - \alpha x$ for $x > 0$. ►

A SHARPENED VERSION OF TALAGRAND'S THEOREM *Suppose X is drawn from the product measure $\mathbf{P} = \mu^{\otimes n}$ on $\Omega = \mathcal{X}^n$, \mathbb{A} is a subset of Ω of strictly positive probability, and $\alpha \geq 0$. Then $\mathbf{E} \exp r_{\alpha}(X, \mathbb{A}) \leq \mathbf{P}(\mathbb{A})^{-\alpha}$ and, a fortiori, for every $t \geq 0$, we have*

$$\mathbf{P}\{\rho_0(X, \mathbb{A}) \geq t\} \leq \frac{1}{\mathbf{P}(\mathbb{A})^{\alpha}} \exp \left(\frac{-\alpha t^2}{2(\alpha+1)} \right).$$

COROLLARY If $t \geq \sqrt{-2 \log P(\mathbb{A})}$ the Gromov–Milman concentration function with respect to convex distance is bounded by $\alpha(t + \sqrt{2 \log 2}; \rho_0, P) \leq 2e^{-t^2/2}$ and

$$P\{\rho_0(X, \mathbb{A}) \geq t\} \leq \exp\left\{-\frac{1}{2}\left(t - \sqrt{-2 \log P(\mathbb{A})}\right)^2\right\}.$$

PROOF: In view of the final conclusion of Lemma 1, we have

$$\begin{aligned} P\{\rho_0(X, \mathbb{A}) \geq t\} &= P\{\rho_0(X, \mathbb{A})^2 \geq t^2\} \leq P\left\{r_\alpha(X, \mathbb{A}) \geq \frac{\alpha}{2(\alpha+1)}t^2\right\} \\ &\leq \exp\left(\frac{-\alpha t^2}{2(\alpha+1)}\right) E \exp r_\alpha(X, \mathbb{A}) \leq \frac{1}{P(\mathbb{A})^\alpha} \exp\left(\frac{-\alpha t^2}{2(\alpha+1)}\right). \end{aligned}$$

Optimising over $\alpha \geq 0$ in the upper bound yields $\alpha = -1 + t/\sqrt{-2 \log P(\mathbb{A})}$. ▶

The bound we have obtained in the corollary matches the Gaussian tail bound not only qualitatively but in the exponent! As de Moivre’s theorem tells us that the Gaussian bound actually captures the right asymptotic picture for Bernoulli trials, we can now declare victory in the general setting and withdraw with the asymptotic spoils. I invite the reader to marvel as I do at how the elementary induction method has managed to tease out such an exquisitely nuanced picture.

7 The bin-packing problem

Talagrand’s formulation is extraordinarily nimble and it is hard to do justice to the wealth of directions to which it can be applied—it is perhaps the case that the ideas are too recent for a “typical” application to be identified from the stew of varied applications of the principle that have bubbled forth. While there is hence an inescapable element of personal preference in the choice of illustrative applications, those that I have chosen to showcase the use of convex distance are not only famous problems which heretofore have yielded their secrets only through deep, specialised analyses, but from our viewpoint have the great advantage of needing very little background by way of preparation.

In broad brush strokes, in applications of Talagrand’s theorem one tries to find a Lipschitz-style condition on a function, the Gromov–Milman formulation then providing a validation of the slogan of Section 3. In typical applications it is usually easier to find a condition involving the r -Hamming distance; Theorem 4.1 provides the bridge to convex distance. As an illustration we begin with a problem in the optimal allocation of resources in operations research.

Suppose X_1, \dots, X_n is a sequence of bounded random variables drawn by independent sampling from some distribution F with support in the unit interval. We think of the X_j as the sizes of some commodity. Given bins of unit size the objective is to pack the X_j into the bins so as to use the *smallest* number of bins; we are not allowed to divide a given commodity X_j across bins and the

sizes of the commodities packed into a given bin must sum to no more than 1. Let $B(X_1, \dots, X_n)$ be the smallest number of bins needed to pack X_1, \dots, X_n ; this is the *binning number* of the sequence. This is a very complicated, non-linear function. But we may note that the map $(X_1, \dots, X_n) \mapsto B(X_1, \dots, X_n)$ depends honestly on all the variables (indeed it is invariant with respect to permutation of the members of the sequence) and so by our slogan it must be a constant.

We seek a Lipschitz-style condition on the binning number to put this intuition on a stronger footing. Passing to vector notation, if $\mathbf{x} = (x_j, 1 \leq j \leq n)$ is any sequence and K is any subset of the indices $\{1, \dots, n\}$, we introduce the nonce notation $\mathbf{x}_K = (x_j, j \in K)$ for the subsequence picked out by K , $B(\mathbf{x})$ and $B(\mathbf{x}_K)$ denoting the binning numbers of the original sequence \mathbf{x} and the subsequence \mathbf{x}_K , respectively. Given any sequence \mathbf{x} , since we can always combine any two bins that are less than half full and thereby reduce the total number of bins in use, $B(\mathbf{x})$ will account for no more than one bin which is filled to less than half its capacity. It follows that there are at least $B(\mathbf{x}) - 1$ bins each of which is filled to more than half its capacity, whence $\frac{1}{2}[B(\mathbf{x}) - 1] \leq |\mathbf{x}| = \sum_{j=1}^n x_j$. As a basic result, we hence have

LEMMA 1 *The inequality $B(\mathbf{x}) \leq 2 \sum_{j=1}^n x_j + 1$ holds for each $\mathbf{x} \in [0, 1]^n$.*

Now consider two sequences $\mathbf{x} = (x_j, 1 \leq j \leq n)$ and $\mathbf{y} = (y_j, 1 \leq j \leq n)$ representing points in the unit cube $[0, 1]^n$. To eschew trivialities we may suppose $\mathbf{x} \neq \mathbf{0}$ as in this case $B(\mathbf{x}) = 0$ and the Lipschitz-style conditions on $B(\mathbf{x})$ that follow hold trivially. Write $J = \{j : x_j \neq y_j\}$ for the set of indices on which \mathbf{x} and \mathbf{y} differ. Then, for indices $j \in J^c$, the subsequence \mathbf{x}_{J^c} can be accommodated by no more than $B(\mathbf{y}_{J^c}) \leq B(\mathbf{y})$ bins. Packing the subsequences \mathbf{x}_J and \mathbf{x}_{J^c} separately can only increase the inefficiency of the packing and so we have $B(\mathbf{x}) \leq B(\mathbf{x}_{J^c}) + B(\mathbf{x}_J) = B(\mathbf{y}_{J^c}) + B(\mathbf{x}_J) \leq B(\mathbf{y}) + B(\mathbf{x}_J)$. Our lemma keeps the inequality ticking in the right direction by providing a bound on $B(\mathbf{x}_J)$ but a little notational preparation makes matters transparent. Introduce the normalised vector $\mathbf{r} = (r_1, \dots, r_n)$ where $r_j = x_j/\|\mathbf{x}\|$ for each j ; clearly, $\mathbf{r} \geq 0$ and $\|\mathbf{r}\| = 1$. With $h(x_j, y_j) = 1\{x_j \neq y_j\}$ in our usual notation denoting the components of the Hamming error vector $\mathbf{h}(\mathbf{x}, \mathbf{y})$, we then obtain

$$\begin{aligned} B(\mathbf{x}) &\leq B(\mathbf{y}) + B(\mathbf{x}_J) \leq B(\mathbf{y}) + 2 \sum_{j \in J} x_j + 1 = B(\mathbf{y}) + 2\|\mathbf{x}\| \sum_{j \in J} \frac{x_j}{\|\mathbf{x}\|} + 1 \\ &= B(\mathbf{y}) + 2\|\mathbf{x}\| \sum_{j=1}^n r_j h(x_j, y_j) + 1 = B(\mathbf{y}) + 2\|\mathbf{x}\| \rho_{\mathbf{r}}(\mathbf{x}, \mathbf{y}) + 1 \end{aligned}$$

and an, admittedly one-sided, Lipschitz-style condition for B has emerged in terms of \mathbf{r} -Hamming distance. The reader should bear in mind, however, that the selection of weights $\mathbf{r} = \mathbf{r}(\mathbf{x})$ is determined by the \mathbf{x} in view so that the chosen Hamming metric *varies with \mathbf{x}* . For each positive integer m , let $\mathbb{A}_m =$

$\{\mathbf{y} : B(\mathbf{y}) \leq m\}$ be the set of sequences whose binning number does not exceed m . By taking the infimum of both sides as \mathbf{y} is allowed to vary across \mathbb{A}_m , we obtain

$$B(\mathbf{x}) \leq \inf\{B(\mathbf{y}) + 2\|\mathbf{x}\|\rho_r(\mathbf{x}, \mathbf{y}) + 1 : \mathbf{y} \in \mathbb{A}_m\} \leq m + 2\|\mathbf{x}\|\rho_r(\mathbf{x}, \mathbb{A}_m) + 1.$$

As $\rho_{r(\mathbf{x})}(\mathbf{x}, \mathbb{A}_m) \leq \max\{\rho_s(\mathbf{x}, \mathbb{A}_m) : s \geq 0, \|s\| = 1\}$, by Theorem 4.1 we may relate the Lipschitz-style bound to the convex distance.

LEMMA 2 Suppose $m \geq 1$. Then $B(\mathbf{x}) \leq m + 2\|\mathbf{x}\|\rho_0(\mathbf{x}, \mathbb{A}_m) + 1$.

Now for some probability. Suppose $\mathbf{X} = (X_1, \dots, X_n)$ is a random point drawn from the product distribution $F^{\otimes n}$ with support in the unit cube. In order to use Lemma 2 we need some information on the likely size of $\|\mathbf{X}\|$. Now $\|\mathbf{X}\|^2 = X_1^2 + \dots + X_n^2$ is concentrated at $E(\|\mathbf{X}\|^2) = nE(X_1^2)$ by the law of large numbers. As $E(\|\mathbf{X}\|) \leq \sqrt{E(\|\mathbf{X}\|^2)}$ (the Cauchy–Schwarz inequality!), it is hence unlikely that $\|\mathbf{X}\|$ exhibits excursions much beyond $\sqrt{nE(X_1^2)}$. Indeed, by grouping terms, we may write

$$\mathbf{P}\left\{\|\mathbf{X}\| \geq \sqrt{2nE(X_1^2)}\right\} = \mathbf{P}\left\{\|\mathbf{X}\|^2 \geq 2nE(X_1^2)\right\} = \mathbf{P}\left\{\sum_{j=1}^n (X_j^2 - E(X_j^2)) \geq nE(X_1^2)\right\}.$$

As $-a = -E(X_1^2) \leq X_1^2 - E(X_1^2) \leq 1 - E(X_1^2) = 1 - a$, the summands are independent, zero-mean random variables bounded in the interval $[-a, 1 - a]$. Hoeffding’s inequality mops up and we obtain

$$\mathbf{P}\left\{\|\mathbf{X}\| \geq \sqrt{2nE(X_1^2)}\right\} \leq e^{-2nE(X_1^2)}. \quad (7.1)$$

Now the occurrence of the event $\{B(\mathbf{X}) \geq m + 2t\sqrt{2nE(X_1^2)} + 1\}$ implies the occurrence of the event $\{\|\mathbf{X}\| \geq \sqrt{2nE(X_1^2)}\}$ or the event $\{\rho_0(\mathbf{X}, \mathbb{A}_m) \geq t\}$. By Boole’s inequality, it follows that

$$\mathbf{P}\left\{B(\mathbf{X}) \geq m + 2t\sqrt{2nE(X_1^2)} + 1\right\} \leq \mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A}_m) \geq t\} + \mathbf{P}\left\{\|\mathbf{X}\| \geq \sqrt{2nE(X_1^2)}\right\}.$$

The second term on the right may be bounded by (7.1) while the first term has the bound $\mathbf{P}\{\rho_0(\mathbf{X}, \mathbb{A}_m) \geq t\} \leq e^{-t^2/4}/\mathbf{P}\{B(\mathbf{X}) \leq m\}$ by Talagrand’s inequality. Clearing the denominator on the right by multiplying throughout by $\mathbf{P}\{B(\mathbf{X}) \leq m\} \leq 1$ we obtain the “symmetrised” tail bound

$$\mathbf{P}\{B(\mathbf{X}) \leq m\} \cdot \mathbf{P}\left\{B(\mathbf{X}) \geq m + 2t\sqrt{2nE(X_1^2)} + 1\right\} \leq e^{-t^2/4} + e^{-2nE(X_1^2)}.$$

Consolidating the terms of the upper bound by setting $u = 2t\sqrt{2nE(X_1^2)}$ and $t \leq \sqrt{8nE(X_1^2)}$, we obtain a characteristic inequality for the binning number.

LEMMA 3 *If $0 \leq u \leq 8n E(X_1^2)^{3/2}$ then*

$$P\{B(\mathbf{X}) \leq m\} \cdot P\{B(\mathbf{X}) \geq m + u + 1\} \leq 2 \exp\left(\frac{-u^2}{32n E(X_1^2)}\right). \quad (7.2)$$

While the inequality (7.2) appears to mix the two tails, we may consolidate by focusing on the median. Suppose that M is any median of $B(\mathbf{X})$: $P\{B(\mathbf{X}) \geq M\} \geq 1/2$ and $P\{B(\mathbf{X}) \leq M\} \geq 1/2$. Setting $m = M$ and $m = M - u - 1$ in turn in (7.2) we see then that both tails, $P\{B(\mathbf{X}) \geq M + u + 1\}$ and $P\{B(\mathbf{X}) \leq M - u - 1\}$, are bounded above by $4 \exp\{-u^2/(32n E(X_1^2))\}$.

THEOREM *Let M be any median of $B(\mathbf{X})$ and suppose $0 \leq u \leq 8n E(X_1^2)^{3/2}$. Then*

$$P\{|B(\mathbf{X}) - M| \geq u + 1\} \leq 8 \exp\left(\frac{-u^2}{32n E(X_1^2)}\right).$$

We have very strong concentration indeed—within the order of $\sqrt{n E(X_1^2)}$ from the median. I have made no particular efforts to find the best constants and the reader who wishes to improve the range of the result should revisit (7.1) and write down a tail bound for, say, $P\{\|\mathbf{X}\| \geq \sqrt{n E(X_1^2) + s}\}$.

The reader will also remark that I have made no effort to estimate the median (or even the expectation). Such calculations can be involved in their own right, depending both on the nature of the distribution assumed and the structure of the problem. But in many applications there is value in knowing that there is concentration even with the point of concentration unspecified. In this and the following sections I shall take this view and resolutely avoid calculating expectations in particular cases. The reader who is interested will find sample calculations in the *Problems*.

8 The longest increasing subsequence

This is a celebrated problem with a rich history. Suppose $\mathbf{x} = (x_1, \dots, x_n)$ is a finite sequence of real numbers. We say that a subsequence $(x_{i_1}, \dots, x_{i_k})$ is increasing if $x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_k}$; if $K = \{i_1, \dots, i_k\}$ we say that K is a *witness* for an increasing subsequence. Suppose now that the random sequence $\mathbf{X} = (X_1, \dots, X_n)$ is drawn by independent sampling from some probability measure F on the real line. Let the random variable $L(\mathbf{X})$ denote the length of the longest increasing subsequence of \mathbf{X} . There is no easy characterisation of $L(\mathbf{X})$ but we note that L is an honest function whence by our slogan $L(\mathbf{X})$ must be essentially constant!

In order to formulate a Lipschitz-style condition we will need to examine the structure of this function more carefully. Consider any sequence $\mathbf{x} = (x_1, \dots, x_n)$. Suppose $L(\mathbf{x}) = k$ and let K be any subset of k indices

$i_1 \leq i_2 \leq \dots \leq i_k$ for which $x_{i_1} \leq x_{i_2} \leq \dots \leq x_{i_k}$, that is to say, K is a witness for a maximal increasing subsequence. Then $L(x) = \text{card } K$. Let $y = (y_1, \dots, y_n)$ be any other sequence. Any subset of K on which the components of x and y agree certainly is a witness for an increasing subsequence of y . Accordingly,

$$L(y) \geq \text{card}\{j \in K : y_j = x_j\} = \text{card } K - \text{card}\{j \in K : y_j \neq x_j\} = L(x) - \sum_{j \in K} h(x_j, y_j).$$

As in the bin-packing problem we are led to look for a Hamming metric. Introduce the Hamming weight vector $r = (r_1, \dots, r_n)$ which puts equal mass $(\text{card } K)^{-1/2} = L(x)^{-1/2}$ precisely on the indices in K , that is, $r_j = L(x)^{-1/2}$ if $j \in K$ and $r_j = 0$ otherwise. It is clear that r is properly normalised with $r \geq 0$ and $\|r\| = 1$. Rewriting our basic inequality in terms of r we obtain

$$L(y) \geq L(x) - \sqrt{L(x)} \sum_{j=1}^n r_j h(x_j, y_j) = L(x) - \sqrt{L(x)} \rho_r(x, y). \quad (8.1)$$

We again have a roving metric with $r = r(x)$ determined by x .

For each $m \geq 1$, let $\mathbb{A}_m = \{y : L(y) \leq m\}$ be the set of sequences in \mathbb{R}^n the length of whose longest increasing subsequence is not larger than m . Then (8.1) implies that $L(x) \leq m + \sqrt{L(x)} \rho_{r(x)}(x, y)$ for all $y \in \mathbb{A}_m$. Taking the infimum of both sides as y is allowed to vary across \mathbb{A}_m yields a result very much in the spirit of Lemma 7.2.

LEMMA Suppose $m \geq 1$. Then, for every sequence $x \in \mathbb{R}^n$, we have

$$L(x) \leq m + \sqrt{L(x)} \rho_0(x, \mathbb{A}_m). \quad (8.2)$$

THEOREM Suppose M is any median of $L(X)$ and $t \geq 0$. Then

$$\mathbf{P}\{L(X) \geq M+t\} \leq 2 \exp\left(\frac{-t^2}{4(M+t)}\right) \text{ and } \mathbf{P}\{L(X) \leq M-t\} \leq 2 \exp\left(\frac{-t^2}{4M}\right). \quad (8.3)$$

PROOF: Rewriting (8.2) in the form $\rho_0(x, \mathbb{A}_m) \geq \sqrt{L(x)} - m/\sqrt{L(x)}$ we are led to consider the function $g(u) = \sqrt{u} - m/\sqrt{u}$ on the positive half-line $(0, \infty)$. Differentiation shows that $g'(u) = \frac{1}{2}(u+m)u^{-3/2} > 0$ for every choice of $m \geq 0$ and so g is an increasing function of its argument.

Now consider the set $\mathbb{B}_{m+t} = \{x : L(x) \geq m+t\}$ comprised of those sequences for which the length of the longest subsequence is at least $m+t$. If $x \in \mathbb{B}_{m+t}$ then $\rho_0(x, \mathbb{A}_m) \geq g(L(x)) \geq g(m+t) = t/\sqrt{m+t}$. By Talagrand's inequality it follows that

$$\mathbf{P}\{X \in \mathbb{B}_{m+t}\} \leq \mathbf{P}\left\{\rho_0(X, \mathbb{A}_m) \geq \frac{t}{\sqrt{m+t}}\right\} \leq \frac{1}{\mathbf{P}(X \in \mathbb{A}_m)} \exp\left(\frac{-t^2}{4(m+t)}\right).$$

Clearing the denominator on the right and expressing the probabilities more intelligibly in terms of the tails of $L(\cdot)$, we obtain

$$P\{L(X) \leq m\} \cdot P\{L(X) \geq m+t\} \leq \exp\left(\frac{-t^2}{4(m+t)}\right).$$

Setting $m = M$ and $m = M - t$ in turn yields the two parts of (8.3). ▶

More detailed investigations show that the concentration inequalities don't tell the whole story: they provide a qualitatively correct picture of the upper tail but do not quite catch the lower-tail behaviour which is sub-Gaussian. But to obtain results of such precision with so little effort is impressive.

9 Hilbert fills space with a curve

Our next illustration of concentration is in the setting of the archetypal hard problem in scheduling. We begin with a delicious historical detour as preparation.

Let $f: [0, 1] \rightarrow [0, 1]^2$ be any function from the closed unit interval into the closed unit square. If f is continuous then as t varies from 0 to 1 the values $f(t)$ trace out a continuous, unbroken curve in the unit square. We can codify the mental picture by dealing with the *image* under f of the unit interval. If A is any subset of $[0, 1]$ write $f_*A = \{f(t) : t \in A\}$ for the image under f of A in the unit square. If f is continuous then we say that the image $f_*[0, 1]$ of the unit interval under f describes a *curve* inside the unit square; we say, naturally enough, that the curve *begins* at $f(0)$ and *ends* at $f(1)$. In a harmless abuse of terminology we shall sometimes identify f with the curve f_* it engenders.

A curve $f_*[0, 1]$ is a one-dimensional line (length without breadth) and so it appears clear that it must occupy a negligible proportion of the area of the square in two dimensions. The geometrical intuition is very strong here and it came as a nasty jar to mathematicians therefore when in 1890 Giuseppe Peano⁸ constructed a curve which filled the entire square.

THEOREM 1 *There exists a space-filling curve, that is to say, a continuous function $f: [0, 1] \rightarrow [0, 1]^2$ with $f_*[0, 1] = [0, 1]^2$.*

Peano's construction was algebraic (though he must have had a very clear geometric picture in mind) but in the following year David Hilbert⁹ provided another example which, borrowing from E. H. Moore, is luminous to the geometric imagination. This is the construction I shall describe.

⁸G. Peano, "Sur une courbe qui remplit toute une aire plane", *Mathematische Annalen*, vol. 36, pp. 157–160, 1890.

⁹D. Hilbert, "Über die stetige Abbildung einer Linie auf ein Flächenstück", *Mathematische Annalen*, vol. 38, pp. 459–460, 1891.

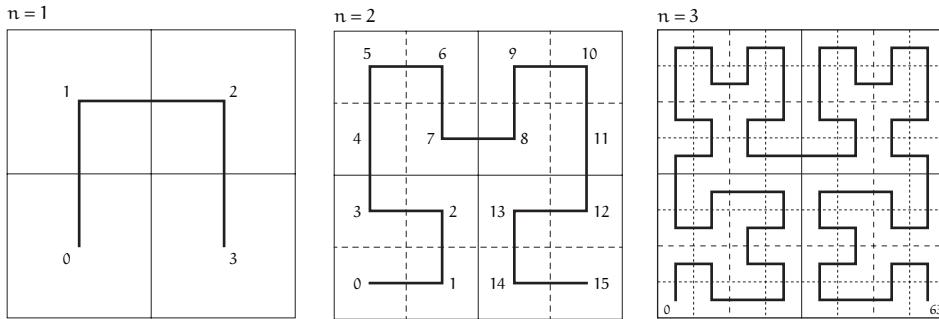


Figure 3: The first three stages in the construction of the Hilbert curve.

The idea is to proceed in stages $n = 0, 1, \dots$, at each stage recursively partitioning the closed unit interval $[0, 1]$ into congruent replicas $\{\mathbb{E}_{nj}, 0 \leq j \leq 4^n - 1\}$ of length 4^{-n} apiece and the unit square $[0, 1]^2$ into congruent replicas $\{\mathbb{F}_{nj}, 0 \leq j \leq 4^n - 1\}$ of side 2^{-n} apiece, associating subintervals with subsquares $\mathbb{E}_{nj} \leftrightarrow \mathbb{F}_{nj}$ in such a way that adjacency relations and inclusion relations are preserved. Here is the procedure. Begin with the initial stage $n = 0$ by associating the closed unit interval $\mathbb{E}_{00} = [0, 1]$ with the closed unit square $\mathbb{F}_{00} = [0, 1]^2$. In the first stage $n = 1$ we partition the unit interval into four contiguous, congruent, and closed subintervals, each of length 4^{-1} , and associate them with four congruent subsquares, each of side 2^{-1} , in the order shown. In the next stage, $n = 2$, we partition each subinterval into four congruent subintervals, each of length 4^{-2} , and each subsquare into four congruent subsquares, each of side 2^{-2} , and use the stage $n = 1$ as a template to stitch together a contiguous sequence of associations as shown in Figures 3. We now proceed recursively, at each stage partitioning subintervals into congruent fourths and subsquares into congruent fourths, and repeat the basic procedure using the previous stage as a template to create a connected pastiche. Figure 4 illustrates the fourth stage, $n = 4$. The procedure preserves two basic properties for every n :

- ① *Contiguity.* The subsquares \mathbb{F}_{nj} and $\mathbb{F}_{n,j+1}$ associated with adjacent subintervals \mathbb{E}_{nj} and $\mathbb{E}_{n,j+1}$ have an edge in common: *adjacency is preserved*.
- ② *Consanguinity.* All subsquares of \mathbb{F}_{nj} are associated with the subintervals of \mathbb{E}_{nj} : *inclusion relations are preserved*.

If the reader is not convinced by the geometry she is invited to proceed by induction.

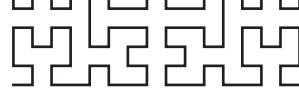


Figure 4: The fourth stage.

Let's take stock. To each point t in the closed unit interval $[0, 1]$ we may associate a nested sequence of subintervals $\mathbb{E}_{00} \supset \mathbb{E}_{1j_1} \supset \mathbb{E}_{2j_2} \supset \dots \supset \mathbb{E}_{nj_n}$ each of which contains t . As the lengths of these subintervals shrink to zero it follows perforce that $\bigcap_n \mathbb{E}_{nj_n}$ consists of the single point t . (If t has a terminating quaternary expansion then there are two such nested sequences; no matter, pick one.) The associated subsquare sequence is also nested, $\mathbb{F}_{00} \supset \mathbb{F}_{1j_1} \supset \dots \supset \mathbb{F}_{nj_n} \supset \dots$ with diagonals shrinking to a point $x = (x_1, x_2)$. (Because adjacent subintervals are mapped into adjacent subsquares, if t has a terminating quaternary expansion both nested sequences containing it lead to the same image x .) Hilbert's recursive procedure hence determines a map $f^H: t \mapsto x$ taking the unit interval $[0, 1]$ into the unit square $[0, 1]^2$. With candidate function in hand we proceed to show that it has the requisite properties.

THEOREM 2 *The map $f^H: [0, 1] \rightarrow [0, 1]^2$ is a surjection, that is to say, for each $x \in [0, 1]^2$ there exists $t \in [0, 1]$ such that $f(t) = x$.*

PROOF: Pick any point $x = (x_1, x_2)$ in $[0, 1]^2$. It belongs to a nested sequence of subsquares whose diagonals shrink into the point x . The corresponding nested sequence of subintervals shrinks into a unique point t in $[0, 1]$. ▶

The theorem does *not* say that the map f^H is injective. Indeed, if x lies on an interior edge (or a corner) of a square then it will belong to two or more subsquares not corresponding to adjacent subintervals. Such points will belong to two or more distinct sequences of nested subsquares and hence will have multiple preimages t_1, t_2, \dots in the closed unit interval all of which are mapped into x .

We will finish off the proof of Theorem 1 if we can show that f^H is continuous. We shall show a little more besides. Identify $\|\cdot\|$ with the usual Euclidean norm in two-space.

THEOREM 3 *For any $s, t \in [0, 1]$, we have $\|f^H(s) - f^H(t)\| < 2\sqrt{5}|s - t|^{1/2}$. In other words, f^H is Lipschitz of order $1/2$ and a fortiori is continuous.*

PROOF: Pick any $s, t \in [0, 1]$. Then there exists a unique $n \geq 0$ such that $4^{-n-1} < |s - t| \leq 4^{-n}$. Thus, either both s and t belong to a common subinterval \mathbb{E}_{nj} or lie in adjacent subintervals \mathbb{E}_{nj} and $\mathbb{E}_{n,j+1}$. Their images $f^H(s)$ and $f^H(t)$, respectively, will either both be in the subsquare \mathbb{F}_{nj} of diagonal $\sqrt{2} \cdot 2^{-n}$ or will lie in adjacent subsquares \mathbb{F}_{nj} and $\mathbb{F}_{n,j+1}$. In the latter case the two subsquares make a rectangle whose diagonal is $\sqrt{5} \cdot 2^{-n}$. Thus, in all cases, $\|f^H(s) - f^H(t)\| \leq \sqrt{5} \cdot 2^{-n} < 2\sqrt{5}|s - t|^{1/2}$. ▶

We call $f_*^H[0, 1]$ the *Hilbert curve*. Our construction has given us an explicit advertisement for Theorem 1.

COROLLARY *The Hilbert curve is space-filling: $f_*^H[0, 1] = [0, 1]^2$.*

We may as well gather some more low-hanging fruit. The recursive procedure shows that the Hilbert curve is *very* wiggly. The reader who has read Section X.8 may find a shrewd suspicion come to mind.

THEOREM 4 *The coordinate functions of the Hilbert curve f^H are continuous everywhere and differentiable nowhere.*

PROOF: Fix any $t \in [0, 1]$. Then for every $n \geq 2$ we may select a point t_n satisfying $9 \cdot 4^{-n} < |t - t_n| \leq 10 \cdot 4^{-n}$. A glance at Figure 3 (or induction, if you must) shows that $f^H(t)$ and $f^H(t_n)$ are separated in both coordinates by at least one square of side 2^{-n} . Writing $f^H = (f_1^H, f_2^H)$ explicitly in terms of its coordinate functions, we see then that

$$\frac{|f_1^H(t) - f_1^H(t_n)|}{|t - t_n|} \geq \frac{2^{-n}}{10 \cdot 4^{-n}} = \frac{2^n}{10},$$

with an identical bound for the second coordinate function. It follows that neither of the coordinate functions has a derivative anywhere. ▶

The reader who is partial to analytical proofs (and objects to the geometric proofs provided here) may wish to try her hand at a recursive analytical specification of Hilbert's procedure. She will get forrader if she represents points in the unit interval in a quaternary representation, $t = (.t_1 t_2 \dots)_4 = \sum_{k \geq 1} t_k 4^{-k}$ (where each $t_k \in \{0, 1, 2, 3\}$), and recognises that the procedure operates by the repeated application of four group transformations, scale, shift, rotation, and reflection. The reader who would like to explore the subject further will find an engaging read in Hans Sagan's monograph.¹⁰

Space-filling curves were at first greeted with consternation, even the very meaning of "curve" and "dimension" apparently called into question by their existence. The passage of time quieted the more extreme reactions to them and they eventually came to be regarded as curiosities, quaint but of no moment. Interest in them was renewed late in the twentieth century with the identification of fractal domains and with renewed interest came the discovery of applications in varied settings. I shall describe one in the following section.

10 The problem of the travelling salesman

Perhaps the most studied problem in combinatorial optimisation is that of the travelling salesman. It is beguilingly simple to state. Suppose there are n cities scattered in the unit square $[0, 1]^2$ at coordinate locations x_1, \dots, x_n . A *tour*

¹⁰H. Sagan, *Space-filling curves*. New York: Springer-Verlag, 1994.

through these cities is a permutation $\Pi: k \mapsto \Pi_k$ of the indices from 1 to n with the cities visited in the order $\Pi_1 \rightarrow \Pi_2 \rightarrow \dots \rightarrow \Pi_n \rightarrow \Pi_1$. The objective of the salesman is to find a tour that minimises the total travel length

$$\|\mathbf{x}_{\Pi_1} - \mathbf{x}_{\Pi_2}\| + \|\mathbf{x}_{\Pi_2} - \mathbf{x}_{\Pi_3}\| + \dots + \|\mathbf{x}_{\Pi_{n-1}} - \mathbf{x}_{\Pi_n}\| + \|\mathbf{x}_{\Pi_n} - \mathbf{x}_{\Pi_1}\|. \quad (10.1)$$

A brute-force approach to tour selection quickly founders upon the realisation that there are $n!$ permutations to evaluate and, excepting when n is very small, this quickly exceeds the capabilities of even the most powerful digital computers. Remarkably, given its simple statement, no general algorithm is known that will efficiently generate an optimal tour in every instance. Indeed, there is some reason to believe that there is no computationally effective procedure which will work in every instance—the problem is one of a large class of problems known to computer scientists as NP-hard problems which are commonly thought to be intractable (in the worst case). Of course, the lack of a universal algorithm should not inhibit the search for one. And if the algorithm we discover is not provably optimal in every case it is at least better than nothing. Indeed the examination of heuristics for difficult problems of this stripe has led to profound advances in the theory of algorithms.

A charming and unexpected heuristic for the travelling salesman problem was proposed by L. K. Platzman and J. J. Bartholdi in 1989.¹¹ Their algorithm for tour selection used space-filling curves in an essential way. I will specialise their heuristic to the Hilbert curve f^H for definiteness though we could select any other space-filling curve of Lipschitz order 1/2. Given a sequence of city locations, $\mathbf{r} = (\mathbf{x}_j, 1 \leq j \leq n)$, the heuristic proceeds in three steps:

- ① Determine preimages under f^H of $\mathbf{x}_1, \dots, \mathbf{x}_n$, that is to say, select points t_1, \dots, t_n in the unit interval so that $f^H(t_1) = \mathbf{x}_1, \dots, f^H(t_n) = \mathbf{x}_n$.
- ② Reorder the points t_1, \dots, t_n in increasing order, $0 \leq t_{(1)} < \dots < t_{(n)} \leq 1$. These are of course the order statistics of the preimages.
- ③ Specify the *Hilbert space-filling tour* $\Pi = \Pi^H(\mathbf{r})$ by identifying $f^H(t_{(k)}) = \mathbf{x}_{\Pi_k}$ for $1 \leq k \leq n$.

As, by Theorem 9.3, the Hilbert curve f^H is Lipschitz of order 1/2, identifying $n+1$ with 1, we see that

$$\begin{aligned} \sum_{j=1}^n \|\mathbf{x}_{\Pi_j} - \mathbf{x}_{\Pi_{j+1}}\|^2 &= \sum_{j=1}^n \|f^H(t_{(j)}) - f^H(t_{(j+1)})\|^2 \leq 20 \sum_{j=1}^n |t_{(j)} - t_{(j+1)}| \\ &= 20 \sum_{j=1}^{n-1} (t_{(j+1)} - t_{(j)}) + 20(t_{(n)} - t_{(1)}) = 40(t_{(n)} - t_{(1)}) \leq 40. \end{aligned} \quad (10.2)$$

¹¹L. K. Platzman and J. J. Bartholdi, “Space-filling curves and the planar travelling salesman problem”, *Journal of the Association for Computing Machinery*, vol. 36, pp. 719–737, 1989.

The reader may well find this remarkable, the fact well worthy of enshrining.

THEOREM 1 *For any n and any choice of city locations $\mathbf{x} = (x_j, 1 \leq j \leq n)$, the sum of the squares of the inter-city distances along the Hilbert space-filling tour is bounded by an absolute constant.*

We remark in passing that the appearance of the constant 40 in our bound is an artefact due to our use of the Hilbert curve; other choices of space-filling curves will yield different constants, some smaller, a fact that is important in practical considerations. Let us promptly put this interesting fact to good use.

With $\mathbf{x} = (x_j, 1 \leq j \leq n)$, let $L(\mathbf{x})$ denote the length of the optimal tour which minimises the tour length (10.1). Suppose $\mathbf{y} = (y_j, 1 \leq j \leq n)$ is another sequence of city locations in the unit square $[0, 1]^2$ with $L(\mathbf{y})$ the length of its optimal tour. We now set about trying to find a Lipschitz-style condition in the spirit of (8.1) to connect $L(\mathbf{x})$ and $L(\mathbf{y})$.

Let \mathcal{X} and \mathcal{Y} denote the collection of cities corresponding to \mathbf{x} and \mathbf{y} , respectively. Their union, $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, is in general a larger collection of cities corresponding to which are the coordinate locations $\mathbf{z} = (z_j)$ in some order. The key to the analysis is the apparently pedestrian observation that the shortest tour through \mathcal{Z} is at least as long as the shortest tour through \mathcal{X} : $L(\mathbf{x}) \leq L(\mathbf{z})$. (The triangle inequality in Euclidean space!) Now \mathcal{X} and \mathcal{Y} may be disjoint or may share one or more cities. To begin, suppose $\mathcal{X} \cap \mathcal{Y} \neq \emptyset$. Suppose we are somehow given an optimal tour for the collection \mathcal{Y} . While it may be hard to find an optimal tour for the original collection \mathcal{X} we can certainly equip it with a Hilbert space-filling tour and we do so. Illustrative tours for \mathcal{Y} and \mathcal{X} are shown in Figure 5. The black bullets labelled A through E represent cities that

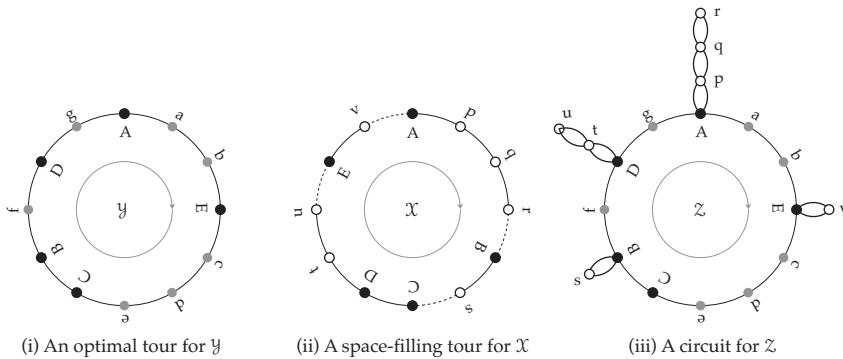


Figure 5: Creating a circuit for $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$ by grafting segments from $\mathcal{X} \setminus \mathcal{Y}$ onto a tour for \mathcal{Y} .

are common to \mathcal{X} and \mathcal{Y} . The bullets shaded grey and labelled a through g in the optimal tour for \mathcal{Y} represent cities in $\mathcal{Y} \setminus \mathcal{X}$; likewise, the white bullets labelled p through v in the Hilbert space-filling tour for \mathcal{X} represent cities in $\mathcal{X} \setminus \mathcal{Y}$. As we

proceed along the space-filling tour for \mathcal{X} we encounter strings of cities in $\mathcal{X} \setminus \mathcal{Y}$ punctuated by cities in $\mathcal{X} \cap \mathcal{Y}$. Call each maximal unbroken sequence of cities in $\mathcal{X} \setminus \mathcal{Y}$ encountered in this tour a *segment*. There are four segments in the tour shown: $(p \rightarrow q \rightarrow r)$, (s) , $(t \rightarrow u)$, and (v) .

In order to form a circuit through the cities in $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$ the idea now is to join forces and use the optimal tour for \mathcal{Y} where possible and the Hilbert space-filling tour for \mathcal{X} where not. The vague but profitable principle that is put to work here is that we should expend effort on those cities in \mathcal{X} that are hard to accommodate, in this case the cities in $\mathcal{X} \setminus \mathcal{Y}$. So, we begin by proceeding along the optimal tour for \mathcal{Y} until we encounter a city in $\mathcal{X} \cap \mathcal{Y}$. We then briefly detour from the optimal tour for \mathcal{Y} by following the corresponding segment (if any) leading from this city in the Hilbert space-filling tour for \mathcal{X} till we encounter the final city in the segment and then simply retrace the path along the segment in reverse to the originating city in $\mathcal{X} \cap \mathcal{Y}$. We then resume the optimal tour for \mathcal{Y} , repeating the process of adding a segment to the circuit each time a city in $\mathcal{X} \cap \mathcal{Y}$ is encountered, until the tour for \mathcal{Y} is complete. This process visits every city in $\mathcal{X} \setminus \mathcal{Y}$ in addition to the cities of \mathcal{Y} and so is a circuit of $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. What is its length? The circuit consists of the optimal tour for \mathcal{Y} together with cycles which pick up each of the segments of the space-filling tour for \mathcal{X} . For each city in a segment, the cycle through it traverses each of the two edges incident on it no more than twice. Accordingly, with $n + 1$ again identified with 1, let $\beta_j(\mathfrak{x}) = 2(\|\mathbf{x}_{\Pi_{j-1}^H(\mathfrak{x})} - \mathbf{x}_{\Pi_j^H(\mathfrak{x})}\| + \|\mathbf{x}_{\Pi_j^H(\mathfrak{x})} - \mathbf{x}_{\Pi_{j+1}^H(\mathfrak{x})}\|)$ be twice the sum of the lengths of the two edges incident upon the j th city in the Hilbert space-filling tour $\Pi^H(\mathfrak{x})$ for \mathcal{X} . Then the sum of the lengths of the cycles through the segments is certainly no more than $\sum_j \beta_j(\mathfrak{x}) h(\mathbf{x}_{\Pi_j^H(\mathfrak{x})}, \mathbf{y}_{\Pi_j^H(\mathfrak{x})})$. As the circuit we have constructed has length surely at least as large as that of an optimal tour of \mathcal{Z} , it follows that $L(\mathfrak{x}) \leq L(\mathfrak{y}) + \sum_j \beta_j(\mathfrak{x}) h(\mathbf{x}_{\Pi_j^H(\mathfrak{x})}, \mathbf{y}_{\Pi_j^H(\mathfrak{x})})$. We have derived this inequality assuming $\mathcal{X} \cap \mathcal{Y} \neq \emptyset$. But we may trivially extend the inequality to include the situation when $\mathcal{X} \cap \mathcal{Y} = \emptyset$ as, in this case, the sum on the right equals four times the length of the Hilbert space-filling tour of \mathcal{X} and so is surely larger than the length of an optimal tour through \mathcal{X} .

To express the bound in terms of Hamming distance we set $\beta(\mathfrak{x}) = (\beta_1(\mathfrak{x}), \dots, \beta_n(\mathfrak{x}))$ and form the unit vector $\mathbf{r}(\mathfrak{x}) = \beta(\mathfrak{x}) / \|\beta(\mathfrak{x})\|$. Then, for every choice of $\mathfrak{x} = (\mathbf{x}_j, 1 \leq j \leq n)$ and $\mathfrak{y} = (\mathbf{y}_j, 1 \leq j \leq n)$, we see that

$$L(\mathfrak{x}) \leq L(\mathfrak{y}) + \|\beta(\mathfrak{x})\| \rho_{\mathbf{r}(\mathfrak{x})}(\mathfrak{x}, \mathfrak{y}) \quad (10.3)$$

and we have our desired Lipschitz-style bound with respect to a roving metric with weight function $\mathbf{r}(\mathfrak{x})$ determined by \mathfrak{x} . To massage our inequality into a familiar form, write $\mathbb{A}_m = \{\mathfrak{y} : L(\mathfrak{y}) \leq m\}$ for the set of n -city configurations in the unit square for which there exists an optimal tour of length no more than m . Taking the infimum of both sides of (10.3) over $\mathfrak{y} \in \mathbb{A}_m$, we see hence that $L(\mathfrak{x}) \leq m + \|\beta(\mathfrak{x})\| \rho_{\mathbf{r}(\mathfrak{x})}(\mathfrak{x}, \mathbb{A}_m)$ for every $m \geq 0$. We can simplify matters further

by the elementary observation $(a + b)^2 \leq 2(a^2 + b^2)$ (the simplest illustration of the Cauchy–Schwarz inequality!) so that by (10.2) we have the bound $\|\beta(\mathbf{r})\|^2 \leq 640$. We finish up by bounding $r(\mathbf{r})$ -Hamming distance from above by the convex distance.

LEMMA *Suppose $\mathbf{r} = (x_j, 1 \leq j \leq n)$ is a sequence of city locations in the unit square. Then $L(\mathbf{r}) \leq m + 8\sqrt{10}\rho_0(\mathbf{r}, A_m)$ for every $m \geq 0$.*

To obtain concentration inequalities all we have to do now is follow the well-worn path of the previous applications. Let $\mathfrak{X} = (X_1, \dots, X_n)$ be a sequence of city locations in the plane chosen by independent sampling from a common distribution with support in the unit square. Harnessing Talagrand’s theorem to our lemma we then obtain

$$\mathbf{P}\{L(\mathfrak{X}) \geq m + t\} \leq \mathbf{P}\left\{\rho_0(\mathfrak{X}, A_m) \geq \frac{t}{8\sqrt{10}}\right\} \leq \frac{1}{\mathbf{P}\{L(\mathfrak{X}) \leq m\}} \exp\left(\frac{-t^2}{2560}\right).$$

Let M be a median of $L(\mathfrak{X})$. By selecting $m = M$ and $m = M - t$, we obtain

THEOREM 2 *Suppose M is any median of $L(\mathfrak{X})$ and $t \geq 0$. Then*

$$\mathbf{P}\{|L(\mathfrak{X}) - M| \geq t\} \leq 4 \exp\left(\frac{-t^2}{2560}\right).$$

I have been munificent with the bounding. The reader who wishes to improve the constants can save a factor of four in the denominator of the exponential by dispensing with the multiplicative factor of two in the definition of $\beta_j(\mathbf{r})$. This will only require the small modification to the circuit that each segment be grafted onto that city of the two at the end which is closer to an end city of the segment. We can improve matters further by using a space-filling curve with a smaller coverage constant. But these don’t change the essential story of concentration.

The reader interested in practical implementations of space-filling tour heuristics will have realised that I have skipped some essential details. For the space-filling heuristic to be computationally viable it is necessary to be able to rapidly generate preimages of points on the space-filling curve. This is indeed possible though the details depend upon the choice of curve. The reader who is interested should consult the paper of Platzman and Bartholdi for details.

The space-filling heuristic is quirky, charming, fun—and useful. J. J. Bartholdi reports that it has been used to build a routing system for Meals-on-Wheels in Fulton County, Atlanta, Georgia, to route blood delivery by the American Red Cross to hospitals in the Atlanta metropolitan area, to target a space-based laser for the Strategic Defense Initiative (which goes by the popular sobriquet “Star Wars”), and to control a pen-plotter for the drawing of maps. The heuristic yields tours roughly 25% longer than the expected length of an

optimal tour. Bartholdi provides an assessment of the trade-off in building a tour of 15,112 cities in Germany and reports that the space-filling heuristic took about a second to compute on a cheap laptop and generated a tour which, at the leisurely rate of 600 km per day, takes 147 days as opposed to a tour generated by a heavy-duty optimisation package which takes 110 days. He points out wryly that the choice is between using the heuristic to get a route immediately and travel an extra month, or using a network of 110 processors and spending two months computing the shortest route—to save a month of driving.

11 Problems

1. *Expectation and median.* Suppose X is a random variable with mean μ and median M . If $P\{|X - M| > t\} \leq e^{-\alpha t^2}$ for every $t \geq 0$, show that $|\mu - M| \leq \sqrt{\pi}/\sqrt{\alpha}$.

2. *Total variation distance.* Suppose X and Y are arithmetic random variables, $\mathcal{L}(X) = \{p_k, k \in \mathbb{Z}\}$ and $\mathcal{L}(Y) = \{q_k, k \in \mathbb{Z}\}$ their respective laws (distributions). The total variation distance between $\mathcal{L}(X)$ and $\mathcal{L}(Y)$ is defined by $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{\mathbb{A}} |P\{X \in \mathbb{A}\} - P\{Y \in \mathbb{A}\}|$. Show that $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \frac{1}{2} \sum_k |p_k - q_k|$. [Hint: Let $\mathbb{A} = \{k : p_k \geq q_k\}$ and set $p' = \sum_{k \in \mathbb{A}} p_k$ and $q' = \sum_{k \in \mathbb{A}} q_k$. Let X' and Y' be Bernoulli trials with success probabilities p' and q' , respectively. Show that $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = d_{TV}(\mathcal{L}(X'), \mathcal{L}(Y'))$.]

3. *Continuation, Pinsker's inequality.* Connect the total variation distance to the Kullback–Leibler divergence by showing that $D(\mathcal{L}(X), \mathcal{L}(Y)) \geq 2d_{TV}(\mathcal{L}(X), \mathcal{L}(Y))^2$. Argue hence that (XVI.1.1) follows directly from the pair of equations (1.1,1.1') in this chapter. [Hint: With X' and Y' as in the previous problem, show that $D(\mathcal{L}(X), \mathcal{L}(Y)) \geq D(\mathcal{L}(X'), \mathcal{L}(Y'))$ by use of the log sum inequality (see Problem XIV.31). In conjunction with the previous problem this reduces considerations to the case of Bernoulli trials.]

4. *Normal maxima.* Suppose X_1, X_2, \dots are independent, $X_j \sim \mathcal{N}(0, 1)$ for each j . Let $M_n = \max\{X_1, \dots, X_n\}$. Then $M_n/\sqrt{2 \log n} \xrightarrow{P} 1$. [Hint: Argue that $P\{M_n > (1 + \epsilon)\sqrt{2 \log n}\} = 1 - [1 - \Phi(-(1 + \epsilon)\sqrt{2 \log n})]^n$ and use the normal tail bound (Example 1.1). The other direction decays even faster (Theorem X.1.1).]

5. *Continuation, concentration on the unit sphere.* Let $R_n = \sqrt{X_1^2 + \dots + X_n^2}$. Evaluate $P\{|R_n/\sqrt{n} - 1| \geq \epsilon\}$ for $0 < \epsilon < 1$ and conclude that $R_n/\sqrt{n} \xrightarrow{\text{a.e.}} 1$. Now show this directly by application of the Kolmogorov strong law.

6. *Another tail bound for bounded variables.* Suppose X is bounded, $|X| \leq M$, with zero mean and variance σ^2 . By expanding $e^{\lambda X}$ in a Taylor series show that

$$\mathbb{E}(e^{\lambda X}) \leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}(X^2 \cdot X^{k-2}) \leq e^{\sigma^2 g(\lambda)}$$

where $g(\lambda) = (e^{\lambda M} - 1 - \lambda M)/M^2$. Set $B(u) = 2u^{-2}((1+u)\log(1+u) - u)$ for $u > 0$ and, by optimising the exponential bound, show that $P\{X \geq t\} \leq \exp\left\{-\frac{t^2}{2\sigma^2} B\left(\frac{Mt}{\sigma^2}\right)\right\}$.

7. *Continuation.* Let $\zeta(u) = 2(1 + \frac{u}{3})((1+u)\log(1+u) - u)$ for $u > -1$. Verify the inequality $\log(1+u) \geq u/(1+u)$ [which is just another way of writing $1 + x \log x \geq x$

for $x > 0$] and hence show that $\zeta(0) = \zeta'(0) = 0$ and $\zeta''(u) \geq 2$. Conclude by two applications of l'Hôpital's rule that $(1 + \frac{u}{3})B(u) \geq 1$ for all $u > 0$.

8. *Continuation, the inequalities of Bennett and Bernstein.* Suppose X_1, \dots, X_n are independent with $|X_j| \leq M$, $E(X_j) = 0$, and $\text{Var}(X_j) = \sigma_j^2$. Let $S_n = X_1 + \dots + X_n$ and suppose $s^2 \geq \sigma_1^2 + \dots + \sigma_n^2$. Show that

$$P\{|S_n| \geq t\} \leq 2 \exp\left\{\frac{-t^2}{2s^2} B\left(\frac{Mt}{s^2}\right)\right\} \leq 2 \exp\left\{\frac{-t^2}{2(s^2 + Mt/3)}\right\}.$$

The first inequality is that of Bennett; the second of Bernstein. For small deviations t from the mean, these inequalities give estimates very close to the normal tail bound.

9. *Azuma's inequality.* Suppose $0 = X_0, X_1, \dots, X_n$ is a bounded-difference martingale [see Problems XIII.13–56] satisfying $|X_{j+1} - X_j| \leq 1$ for each j . Show that for every $t > 0$ we have $P\{X_n > t\} \leq e^{-t^2/(2^n)}$. Hence show that if $\mu = X_0, X_1, \dots, X_n$ is a martingale with $|X_{j+1} - X_j| \leq 1$ for each j then $P\{|X_n - \mu| > t\} \leq 2e^{-t^2/(2^n)}$. This extension of Hoeffding's inequality is a key component of martingale approaches to concentration. It continues to be of wide general utility though it may be fair to say that to some extent martingale approaches have been superseded by Talagrand's approach to concentration of measure. [Hint: Form the martingale difference $Y_j = X_{j+1} - X_j$ and, writing $X_n = \sum_j Y_j$, condition on X_{n-1} , and proceed as in the proof of Hoeffding's inequality for the binomial in Section XVI.1 to bound $E(e^{\lambda X_n} | X_{n-1})$.]

Efficient communication: How fast can we communicate? A signal vector \mathbf{X} with power-limited components $|X_j| \leq \sqrt{P_s}$ is selected from a code book $\{\mathbf{X}_1, \dots, \mathbf{X}_M\}$ and transmitted over a noisy medium. The received vector $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ is additively corrupted by a Gaussian noise vector \mathbf{Z} whose components are independent $N(0, P_N)$ variables. The optimum receiver maps the noisy received signal \mathbf{Y} into the signal vector in the code book that is closest to it—this is the maximum likelihood receiver. The setting is that of the Gaussian channel of Section XIV.9 in which we discovered an upper bound on the rate of reliable communication. We now propose to analyse what is achievable in the style of Section 2 of this chapter.

10. *Exploiting spherical symmetry.* Suppose s and t are two signal vectors. Given that s , say, is transmitted, argue that the probability that $\mathbf{Y} = s + \mathbf{Z}$ is closer to t than to s is $\Phi(-\|s - t\|/\sqrt{4P_N}) < \prod_{j=1}^n \exp(-(s_j - t_j)^2/8P_N)$.

11. *Signalling with multilevel sequences.* The components X_j of a signal vector \mathbf{X} are selected by independent sampling from the uniform distribution on $\{a_0, a_1, \dots, a_{l-1}\}$ where the $a_i = (-1 + \frac{2j}{l-1})\sqrt{P_s}$ represent l equispaced levels between $-\sqrt{P_s}$ and $+\sqrt{P_s}$. The code book consists of $M = 2^{nR}$ independent signal vectors. Setting

$$R_0 = -\log_2 \left\{ \frac{1}{l^2} \sum_{j=1}^n \sum_{k=1}^l \exp\left(\frac{-(a_j - a_k)^2}{8P_N}\right) \right\},$$

show that the average probability of receiver error is bounded above by $2^{-n(R_0 - R)}$. Conclude that reliable communication is possible at any rate $R < R_0$ bits per dimension. [Hint: Suppose, say, $\mathbf{X} = \mathbf{X}_1$ is selected for transmission. Use Boole's inequality to reduce the setting to pairwise comparisons of two signal vectors \mathbf{X}_1 and \mathbf{X}_j for $j > 1$ and use the previous problem.]

12. *Continuation.* On a semilog scale, plot R_0 versus $10 \log_{10} P_S/P_N$ as $2 \leq l \leq 64$ varies over powers of 2. Plot Shannon's Gaussian channel capacity C given by (XIV.9.4) on the same graph. What do you conclude?

13. *Optimisation over the multilevel constellation.* The uniform selection of signal vectors from the lattice $\{a_0, a_1, \dots, a_{l-1}\}^n$ is qualitatively not optimal as selections in the interior of the lattice place signal vectors in close proximity. Pick each signal vector component instead from a distribution $p = (p_0, p_1, \dots, p_{l-1})$ which places mass p_j at a_j . Setting

$$R_0^* = -\log_2 \left\{ \sum_{j=1}^n \sum_{k=1}^n p_j p_k \exp \left(\frac{-(a_j - a_k)^2}{8P_N} \right) \right\},$$

show that the average probability of receiver error is now bounded by $2^{-n(R_0^* - R)}$ so that reliable communication is possible even at rates up to $R_0^* = R_0^*(p)$ bits per dimension.

14. *Continuation.* Optimise the selection of p by the method of Lagrange multipliers to find the largest rate R_0^* . Repeat Problem 12 with R_0^* replacing R_0 in the ordinate.

The induction method: Talagrand's flexible induction can be applied just as easily to Hamming distance. The setting is that of the product space \mathcal{X}^n equipped with product measure $P = \mu^{\otimes n}$ and Hamming distance ρ_1 or, more generally, r -Hamming distance ρ_r .

15. *Three easy preliminaries.* Define the function $f(p) = p(1 + e^\lambda(1 - p))$ on the unit interval $0 \leq p \leq 1$. Here λ is a positive parameter. Show that $\min_p f(p) = \cosh(\frac{\lambda}{2})^2 \leq e^{\lambda^2/4}$.

16. *Continuation.* If $0 \leq g \leq 1$, show that $\min\{e^\lambda, 1/g\} \leq 1 + e^\lambda(1 - g)$. [Hint: Consider the cases $g > e^{-\lambda}$ and $g \leq e^{-\lambda}$ separately.]

17. *Continuation.* Suppose $g: \mathcal{X} \rightarrow [0, 1]$ is measurable with respect to μ . Using the results of the previous two problems, show that

$$\int_{\mathcal{X}} \min \left\{ e^\lambda, \frac{1}{g(\omega)} \right\} \mu(d\omega) \cdot \int_{\mathcal{X}} g(\omega) \mu(d\omega) \leq e^{\lambda^2/4}.$$

18. *Hamming distance: the base case $n = 1$.* Suppose A is a (measurable) subset of \mathcal{X} of strictly positive probability. Show that $E \exp \lambda \rho_1(X, A) \leq 1 + e^\lambda(1 - P(A))$. Using Problem 15, conclude that $E \exp \lambda \rho_1(X, A) \leq P(A)^{-1} e^{\lambda^2/4}$.

19. *Continuation, the induction step.* Prove by induction that if A is a measurable subset of \mathcal{X}^n of strictly positive probability then $E \exp \lambda \rho_1(X, A) \leq P(A)^{-1} e^{\lambda^2 n/4}$. [Hint: Follow Talagrand's induction method and show that under the induction hypothesis, if A is a (measurable) subset of $\mathcal{X}^n \times \mathcal{X}$, then

$$E \exp \lambda \rho_1((X, \omega), A) = \frac{e^{\lambda^2 n/4}}{P \otimes \mu(A)} \cdot \frac{P \otimes \mu(A)}{P(\Pi A)} \int_{\mathcal{X}} \min \left\{ e^\lambda, \frac{1}{P(\Pi_{\omega} A)/P(\Pi A)} \right\} \mu(d\omega).$$

Now use Problem 17 to complete the induction.]

20. *Continuation, concentration function.* Suppose A is a subset of \mathcal{X}^n and $t \geq 0$. Prove the concentration inequality $P[\rho_1(X, A) \geq t] \leq P(A)^{-1} e^{-t^2/n}$.

21. *Continuation, concentration with respect to Hamming distance.* Suppose $f: \mathcal{X}^n \rightarrow \mathbb{R}$ is Lipschitz (with constant 1) with respect to Hamming distance. Let M_f be any median of f . Then $\mathbf{P}\{|f - M_f| > t\} \leq 4e^{-t^2/n}$. If f has expectation $E_f = E f(\mathbf{X})$ then $\mathbf{P}\{|f - E_f| > t + 2\sqrt{n}\} \leq 4e^{-t^2/n}$. [Hint: See Problem 1.]

22. *Bin-packing.* Show that the binning number satisfies $|B(\mathbf{x}) - B(\mathbf{y})| \leq \rho_1(\mathbf{x}, \mathbf{y})$. If M is any median of $B(\mathbf{X})$, conclude that $\mathbf{P}\{|B(\mathbf{X}) - M| > t\} \leq 4e^{-t^2/n}$. [This is for illustrative purposes only; the convex distance provides much sharper concentration.]

23. *Concentration with respect to r -Hamming distance.* The corresponding results for r -Hamming distance require only that the reader retrace her path through the steps establishing concentration for Hamming distance. Suppose $r \geq 0$, $\lambda \geq 0$, and $t > 0$. Show that $E \exp \lambda \rho_r(\mathbf{X}, \mathbb{A}) \leq \mathbf{P}(\mathbb{A})^{-1} \exp(\lambda^2 \|r\|^2/4)$.

24. *Continuation.* Prove that $\mathbf{P}\{\rho_r(\mathbf{X}, \mathbb{A}) \geq t\} \leq \mathbf{P}(\mathbb{A})^{-1} \exp(-t^2/\|r\|^2)$ and conclude that the concentration function is bounded by $\alpha(t; \rho_r) \leq 2 \exp(-t^2/\|r\|^2)$. In particular, if f is any Lipschitz function (with constant 1) with respect to r -Hamming distance ρ_r on $\Omega = \mathcal{X}^n$ and M_f is any median of f then $\mathbf{P}\{|f - M_f| > t\} \leq 4 \exp(-t^2/\|r\|^2)$ and f is indeed concentrated at its median.

Concentration via convex distance:

25. *The binning number.* Let $\mathbf{P}\{X_1 \leq 1/2\} = p$ in the bin-packing problem. Show that $\max\{n E(X_1), n(1 - p)\} \leq E(B(\mathbf{X})) \leq n(1 - p/2) + 1$.

26. *Space-filling curves.* Show that there exist no space-filling curves of Lipschitz order greater than 1/2.

27. *Random points in the unit square.* Suppose $\mathbf{X}_1, \dots, \mathbf{X}_n$ are selected by independent sampling from the uniform distribution in the unit square $[0, 1]^2$. Let \mathbf{x} be any point in $[0, 1]^2$. Show that $\mathbf{P}\{\min_{1 \leq j \leq n} \|\mathbf{X}_j - \mathbf{x}\| > t\} \leq e^{-\pi n t^2/4}$ and hence that $E(\min_{1 \leq j \leq n} \|\mathbf{X}_j - \mathbf{x}\|) \leq n^{-1/2}$. [Hint: Situate \mathbf{x} at one of the corners to upper bound the probability. Now use Problem XIII.3.]

28. *The travelling salesman.* Suppose $L_n^* = \sup_{x_1, \dots, x_n} L_n(x_1, \dots, x_n)$ denotes the maximal length of an optimal travelling salesman tour through any n -point subset of the unit square $[0, 1]^2$. Show that $\min_{1 \leq i < j \leq n} \|x_i - x_j\| \leq \sqrt{5} n^{-1/2}$ and hence $L_n^* \leq L_{n-1}^* + 2\sqrt{5} n^{-1/2}$. If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are selected by independent sampling from the uniform distribution in $[0, 1]^2$ conclude that $E(L_n(\mathbf{X}_1, \dots, \mathbf{X}_n)) \leq 4\sqrt{5}(\sqrt{n}-1)$. [Hint: Cover the unit square with n congruent subsquares of side $n^{-1/2}$ and use a pigeon-hole argument. Finish up by approximating sums by integrals.]

29. *Continuation.* Show that $E(\min_{1 \leq i < j \leq n} \|\mathbf{X}_i - \mathbf{X}_j\|) \geq c\sqrt{n}$ for some constant c ; hence $E(L_n(\mathbf{X}_1, \dots, \mathbf{X}_n)) \geq c\sqrt{n}$. [It follows that $E(L_n(\mathbf{X}_1, \dots, \mathbf{X}_n))$ has a growth rate exactly of the order of \sqrt{n} . By a more intricate Poissonisation argument it can indeed be shown that there exists a constant $\gamma > 0$ such that $E(L_n(\mathbf{X}_1, \dots, \mathbf{X}_n))/\sqrt{n} \rightarrow \gamma$ and even that $L_n(\mathbf{X}_1, \dots, \mathbf{X}_n))/\sqrt{n} \rightarrow \gamma$ a.e. This is the famous Beardwood–Halton–Hammersley theorem.¹²]

¹²J. Beardwood, J. H. Halton, and J. M. Hammersley, “The shortest path through many points”, *Proceedings of the Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

30. *The free energy in the Sherrington–Kirkpatrick spin glass model.* A vector $\epsilon = (\epsilon_i, 1 \leq i \leq n) \in \{-1, 1\}^n$ represents a configuration of spins which influence each other via a system of interactions $w = (w_{ij}, 1 \leq i < j \leq n) \in \{-1, 1\}^{\binom{n}{2}}$. The partition function is given by $Z_n(w) = 2^{-n} \sum_{\epsilon} \exp(-H_{\epsilon}(w)/T\sqrt{n})$ where T represents temperature and $H_{\epsilon}(w) = \sum_{1 \leq i < j \leq n} w_{ij} \epsilon_i \epsilon_j$ is the free energy of the state ϵ . The quantity $\log Z_n(w)$ represents the free energy per state. Let $r = (r_{ij}, 1 \leq i < j \leq n)$ be the constant unit vector with components $r_{ij} = \binom{n}{2}^{-1}$. If $w, w' \in \{-1, 1\}^{\binom{n}{2}}$, show that $|H_{\epsilon}(w) - H_{\epsilon}(w')| \leq \sqrt{2n(n-1)} \rho_r(w, w')$. Suppose now that the components of w are independent and take values -1 and $+1$ with equal probability $1/2$. Show that

$$P\{\log Z_n(w) \geq m + t\} P\{\log Z_n(w) \leq m\} \leq \exp\left(\frac{-t^2 T^2}{8(n-1)}\right)$$

and hence argue that $\log Z_n(w)$ is concentrated at its median. [In the Sherrington–Kirkpatrick model w is assumed to be a system of independent, standard normal random variables. The binary interactions assumed here simplify calculations.]

31. *Percolation.* Suppose $X = (X_j, 1 \leq j \leq n)$ is a sequence of independent variables and, for each j , there exists r_j such that $r_j \leq X_j \leq r_j + 1$. Suppose $\mathcal{F} \subset \mathbb{R}^n$ is a family of n -tuples $\alpha = (\alpha_j, 1 \leq j \leq n)$ satisfying $\sigma = \sup_{\alpha \in \mathcal{F}} \|\alpha\| < \infty$. Show that $Z_n^{\mathcal{F}} = \sup_{\alpha \in \mathcal{F}} \sum_{j=1}^n \alpha_j X_j$ is concentrated at its median. Variables of this form appear in percolation theory.

32. *Configuration functions.* Motivated by the increasing subsequence problem, say that a positive function $f(x)$ of sequences $x \in \mathbb{R}^n$ is a configuration function if there is a positive function $\psi(u)$ with $\psi''(u) \leq 0$ for $u > 0$ (that is, $-\psi$ is convex) and a constant γ such that $f(x) \leq m + \psi(f(x)) \rho_0(x, \mathbb{A}_m) + \gamma$ for every $m \geq 0$. Here $\mathbb{A}_m = \{y : f(y) \leq m\}$. If f is a configuration function and M any median of f , show that $P\{f(X) \geq M + t + \gamma\} \leq 2 \exp\left(\frac{-t^2}{4\psi(M+t+\gamma)^2}\right)$ and $P\{f(X) \leq M - t - \gamma\} \leq 2 \exp\left(\frac{-t^2}{4\psi(M)^2}\right)$ by mimicking the analysis of the longest increasing subsequence.

33. *The longest common subsequence.* Let $x = (x_1, \dots, x_n)$ and $x' = (x'_1, \dots, x'_n)$ be two sequences. We say that x and x' exhibit a common subsequence of length k if there exists a subset of k indices on which the corresponding subsequences agree. Let $C(x, x')$ be the length of the longest common subsequence of x and x' . Show that $C(x, x')$ is a configuration function (in the sense of the previous problem). If X and X' denote two independent sequences drawn from the same distribution and M is any median of $C(X, X')$, show hence that $P\{C(X, X') \geq M + t\} \leq 2 \exp\left(\frac{-t^2}{8(M+t)}\right)$ and $P\{C(X, X') \leq M - t\} \leq 2 \exp\left(\frac{-t^2}{8M}\right)$. This problem has wide relevance and finds resonance in genetics, speech recognition, and string parsing in computer science.

34. *The independence number of a random graph.* A subset V of vertices of a graph G is independent (in a graph-theoretic sense) if there are no edges of G that connect vertices in V . The independence number of G , denoted $\alpha(G)$, is the size of the largest independent set in G . Show that the independence number is a configuration function (in the sense of Problem 32) and hence that the independence number $\alpha(G_{n,p})$ of an Erdős–Rényi random graph $G_{n,p}$ is concentrated at its median value.

35. *The Vapnik–Chervonenkis dimension.* Show that the Vapnik–Chervonenkis dimension $V_{\mathfrak{A}}(x)$ [see (XVI.12.2)] is a configuration function (in the sense of Problem 32) and hence that $V_{\mathfrak{A}}(X)$ is concentrated at its median.

XVIII

Poisson Approximation

Casual acquaintance with the bell curve has led to an easy acceptance of the folk wisdom that typical or common circumstances are approximated by the normal distribution. Less appreciated is the fact that the Poisson distribution crops up with almost as much frequency when we deal with atypical or uncommon situations.

c 1–5, 7
A 6, 8, 9

The subject of Poisson approximation has a long history. Poisson's approximation to the binomial dates to 1837; the paradigm bearing his name has antecedents in de Montmort's analysis of matchings (*le problème des rencontres*) in 1708 with provenances extending to antiquity in the sieve of Eratosthenes and the totient function. The reader will recall that the paradigm crops up naturally in situations where events are approximately independent. If she has forgotten the details, she should take the opportunity to revisit the sieve methods within which the paradigm appears in Chapter IV.

In a satisfying return to these classical ideas, the last quarter of the twentieth century has seen renewed attention paid to the Poisson paradigm in a shiny new cast and it has come to be realised that the phenomenon is of surprising ubiquity. The following (somewhat inadequate and imprecise) slogan captures the prevailing sentiment.

SLOGAN *Chance-driven situations involving exceedances or extremes tend to follow the Poisson distribution.*

The reader will find persuasive support of the idea espoused in the delightful range of applications showcased by D. Aldous in his monograph.¹

The twentieth-century investigations into the subject were sparked by a powerful idea developed by Charles Stein in 1970 for the approximate computation of expectations.² It was quickly realised that the idea is of potentially wide

¹D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springer-Verlag, 1989.

²C. Stein, *Approximate Computation of Expectations*. Hayward, CA: IMS, 1986.

applicability but the abstract approximation principle is subtle and requires effort to master. Stein's principle takes a particularly elegant form, however, for Poisson approximation and Louis Chen worked out the details in 1975.³ A key feature was the discovery that the Poisson distribution possesses a remarkable property that uniquely characterises it. This discovery exposed a rich vein of hitherto unexploited results in the general theory that is still being actively mined.

The Stein–Chen method provides a framework for the slogan within which the quality of Poisson approximation can be judged. From our perspective it is a happy accident that the main features of the method can be explicated by elementary methods. This chapter deals with this framework and complements the classical probability sieves of Chapter IV.

1 A characterisation of the Poisson

Consider the Poisson distribution $p(k; \lambda) = e^{-\lambda} \lambda^k / k!$ for positive k . The starting point for the development is the apparently innocuous identity

$$kp(k; \lambda) = \lambda p(k - 1; \lambda) \quad (1.1)$$

valid for all k . If \mathbb{K} is any subset of integers, then summing over \mathbb{K} yields an equivalent form

$$\sum_{k \in \mathbb{K}} kp(k; \lambda) = \sum_{k \in \mathbb{K}} \lambda p(k - 1; \lambda). \quad (1.1')$$

Suppose Z represents a Poisson variable of mean λ and let $1_{\mathbb{K}}(k)$ denote the indicator for the set \mathbb{K} . We may then write (1.1') compactly in the form

$$\mathbf{E}(Z 1_{\mathbb{K}}(Z)) = \mathbf{E}(\lambda 1_{\mathbb{K}}(Z + 1)). \quad (1.1'')$$

Chen discovered in 1975 that the representation (1.1'') was in fact uniquely characteristic of the Poisson. We detour to pick up an ancillary result first.

THEOREM 1 *The Poisson distribution is uniquely determined by its moments.*

PROOF: Suppose X is Poisson with mean λ and let $\mu_n = \mu_n(\lambda) = \mathbf{E}(X^n)$. By milking the basic identity (1.1), we obtain

$$\begin{aligned} \mu_n(\lambda) &= \sum_k k^n p(k; \lambda) = \lambda \sum_k k^{n-1} p(k - 1; \lambda) \stackrel{(j=k-1)}{=} \lambda \sum_j (j + 1)^{n-1} p(j; \lambda) \\ &= \lambda \sum_j \sum_k \binom{n-1}{k} j^k p(j; \lambda) = \lambda \sum_k \binom{n-1}{k} \sum_j j^k p(j; \lambda) = \lambda \sum_k \binom{n-1}{k} \mu_k(\lambda). \end{aligned} \quad (1.2)$$

³L. H. Y. Chen, "Poisson approximation for dependent trials", *Annals of Probability*, vol. 3, pp. 534–545, 1975.

By induction, it follows that $\mu_n(\lambda)$ may be expressed as a polynomial of degree n of the form $\lambda^n + c_{n,n-1}\lambda^{n-1} + \cdots + c_{n,2}\lambda^2 + \lambda$ where all the coefficients $c_{n,j}$ are positive. Replacing λ by the maximum of 1 and λ , $\lambda_1 = \max\{1, \lambda\}$, we obtain $\mu_n(\lambda) \leq \lambda_1^n B_n$ where $B_n = \mu_n(1)$ is the n th moment of a Poisson variable of mean 1. Our proof will be complete in view of Theorem XV.1.4 and the inversion theorem for Laplace transforms if we can show that $\frac{1}{n}B_n^{1/n} \rightarrow 0$.

By applying the basic recurrence (1.2) to the case $\lambda = 1$, we obtain the *Bell recurrence* $B_n = \sum_k \binom{n-1}{k} B_k$ with base $B_0 = 1$.⁴ A trick with generating functions allows us to write down an explicit formula for B_n . Define

$$\widehat{B}(s) = \sum_{n=0}^{\infty} B_n \frac{s^n}{n!}. \quad (1.3)$$

This is the “exponential generating function”.⁵ By formal term-by-term differentiation we obtain

$$\begin{aligned} \widehat{B}'(s) &= \sum_{n \geq 1} B_n \frac{s^{n-1}}{(n-1)!} = \sum_{n \geq 1} \sum_k \binom{n-1}{k} B_k \frac{s^{n-1}}{(n-1)!} \\ &= \sum_k B_k \frac{s^k}{k!} \sum_{n \geq k+1} \frac{s^{n-k-1}}{(n-k-1)!} = e^s \widehat{B}(s). \end{aligned}$$

Solving the differential equation shows that $\widehat{B}(s)$ is of the form e^{e^s+c} and as $\widehat{B}(0) = 1$ we obtain $c = -1$. Writing out the power series for the double exponential explicitly, we obtain

$$\widehat{B}(s) = e^{e^s-1} = \frac{1}{e} \sum_{k=0}^{\infty} \frac{e^{ks}}{k!} = \frac{1}{e} \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{k^n s^n}{k! n!} = \sum_{n=0}^{\infty} \frac{s^n}{n!} \left(\frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!} \right). \quad (1.4)$$

By comparing coefficients in the two power series (1.3,1.4) we may now simply read out *Dobinski's formula*

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}. \quad (1.5)$$

While this expression may be used as a starting point for the derivation of precise estimates of B_n ,⁶ the simplest bounds will suffice for our purposes.

⁴While not important for our purposes, the recurrence shows that B_n enumerates the number of partitions of $\{1, \dots, n\}$. It is called the *n*th *Bell number*.

⁵The reader will find a hugely entertaining read on different types of generating functions in H. S. Wilf, *generatingfunctionology*. San Diego: Academic Press, 1990.

⁶See, for instance, D. Berend and T. Tassa, “Improved bounds on Bell numbers and on moments of sums of random variables”, *Probability and Mathematical Statistics*, vol. 30, pp. 185–205, 2010.

The elementary bound $k! \geq k^k e^{-k}$ [see (IV.5.2)] shows that $\log \frac{k^n}{k!} \leq (n - k) \log k + k$. We are accordingly led to consider the function $f_n(x) = (n - x) \log x + x$. Differentiation shows that, for each n , f_n achieves its unique maximum at the point $m = m(n)$ satisfying $m \log m = n$ or $m \sim n/\log n$. In view of the rapid increase of the factorial in the denominator most of the contribution to the sum in Dobinski's formula arises from terms in the immediate vicinity of m . For instance, if $k \geq 2n \geq 56$ we have $f_n(k) \leq f_{k/2}(k) \leq -k$ and so $\sum_{k \geq 2n} e^{f_n(k)} \leq \sum_{k \geq 2n} e^{-k} = e^{-2n}/(1 - e^{-1})$. As the summands $k^n/k!$ on the right in (1.5) are dominated by $e^{f_n(k)}$, we obtain

$$\sum_{k=0}^{\infty} \frac{k^n}{k!} \leq \sum_{k < 2n} e^{f_n(k)} + \sum_{k \geq 2n} e^{f_n(k)} \leq 2ne^{f_n(m)} + \frac{e^{-2n}}{1 - e^{-1}} = 2nm^{n-m}e^m + \frac{e^{-2n}}{1 - e^{-1}}$$

if $n \geq 28$. The contribution from the tail of the series hence dies (at least) exponentially fast. As $m \sim n/\log n$, for all sufficiently large n we obtain $B_n \leq n\left(\frac{n}{\log n}\right)^n$. It follows hence that $\frac{1}{n}B_n^{1/n} = O\left(\frac{1}{\log n}\right) \rightarrow 0$. ▶

We are now ready for Chen's characterisation of the Poisson distribution.

THEOREM 2 Suppose $\lambda > 0$. A positive, arithmetic random variable Z has a Poisson distribution with mean λ if, and only if,

$$E(\lambda g(Z+1) - Zg(Z)) = 0 \quad (1.6)$$

for every bounded function $g: \mathbb{Z}^+ \rightarrow \mathbb{R}$.

PROOF OF NECESSITY: This is the easy part. Suppose Z has a Poisson distribution with mean λ . We then have

$$\begin{aligned} E(\lambda g(Z+1)) &= \sum_k \lambda p(k; \lambda) g(k+1) = \sum_j \lambda p(j-1; \lambda) g(j) \\ &= \sum_j j p(j; \lambda) g(j) = E(Zg(Z)) \end{aligned}$$

by an application of the basic identity (1.1). ▶

PROOF OF SUFFICIENCY: Suppose Z is a positive, arithmetic variable satisfying (1.6) for every bounded g . For every bounded $h: \mathbb{Z}^+ \rightarrow \mathbb{R}$, write $\mu_\lambda(h) = \sum_k p(k; \lambda)h(k)$ for expectation with respect to the Poisson distribution with mean λ . (As usual, E represents expectation with respect to the real underlying distribution of Z .) It will suffice to show that the equation

$$\lambda g(k+1) - kg(k) = h(k) - \mu_\lambda(h) \quad (k \geq 0) \quad (1.7)$$

has a bounded solution g for every choice of bounded h . A simple truncation argument now shows that Z has moments of all orders and, moreover, the moments of Z coincide with those of a Poisson random variable of mean λ . Indeed,

for any $m \geq 0$, form the sequence of positive functions $h_n(k) = k^m 1_{[0,n]}(k)$ which increases pointwise to the limit function $h(k) = k^m$ on the positive integers. Now (1.7) applied to the bounded h_n says that the equation

$$\lambda g_n(k+1) - kg_n(k) = h_n(k) - \mu_\lambda(h_n) \quad (k \geq 0)$$

has a bounded solution g_n . By taking expectation of both sides with respect to the real distribution of Z , we then obtain

$$E(\lambda g_n(Z+1) - Zg_n(Z)) = E h_n(Z) - \mu_\lambda(h_n)$$

and in view of (1.6) this implies $E h_n(Z) = \mu_\lambda(h_n) = \sum_{k=0}^n p(k; \lambda) k^m$. By the monotone convergence theorem we have $E(Z^m 1_{[0,n]}(Z)) = E h_n(Z) \rightarrow E h(Z) = E(Z^m)$ and so $E(Z^m) = \sum_{k=0}^\infty p(k; m) k^m = \mu_\lambda(h)$. Thus Z has the same moments as a $\text{Poisson}(\lambda)$ variable. In view of Theorem 1, the moments of Z determine its distribution and it follows perforce that Z indeed has a Poisson distribution with mean λ . Thus, we only need to show that (1.7) has a bounded solution g . We proceed recursively.

The base of the recurrence may be arbitrarily specified and we may as well simplify matters by setting $g(0) = 0$. We centre the variables with respect to the Poisson distribution by setting $f(k) = h(k) - \mu_\lambda(h)$ whence $\mu_\lambda(f) = \mu_\lambda(h - \mu_\lambda(h)) = 0$. We may now rewrite (1.7) in the form $g(k+1) = \frac{1}{\lambda} f(k) + \frac{k}{\lambda} g(k)$ for $k \geq 0$. Introducing the “falling factorial” notation $k^j = k(k-1)\cdots(k-j+1)$ for $j \geq 0$, a simple induction now shows that

$$g(k+1) = \frac{1}{\lambda} \sum_{j=0}^k \frac{k^j}{\lambda^j} f(k-j) \stackrel{i=k-j}{=} \frac{1}{\lambda} \sum_{i=0}^k \frac{k^{k-i}}{\lambda^{k-i}} f(i) = \frac{k!}{\lambda^{k+1}} \sum_{i=0}^k \frac{\lambda^i}{i!} f(i). \quad (1.8)$$

To show that g is bounded we exploit the fact that f has zero mean with respect to a Poisson averaging. Thus,

$$0 = \mu_\lambda(f) = e^{-\lambda} \sum_{i=0}^\infty \frac{\lambda^i}{i!} f(i) = e^{-\lambda} \sum_{i=0}^k \frac{\lambda^i}{i!} f(i) + e^{-\lambda} \sum_{i=k+1}^\infty \frac{\lambda^i}{i!} f(i).$$

It follows that $\sum_{i=0}^k \lambda^i f(i)/i! = - \sum_{i=k+1}^\infty \lambda^i f(i)/i!$ whence

$$g(k+1) = \frac{-k!}{\lambda^{k+1}} \sum_{i=k+1}^\infty \frac{\lambda^i}{i!} f(i).$$

Suppose $|h| \leq C$. Then $|f| \leq 2C$ and we hence obtain

$$|g(k+1)| \leq 2C \sum_{i=k+1}^\infty \lambda^{i-k-1} \frac{k!}{i!} = 2C \sum_{i=k+1}^\infty \frac{\lambda^{i-k-1}}{i^{i-k-1}} \leq 2C \sum_{i=k+1}^\infty \frac{\lambda^{i-k-1}}{(i-k-1)!} = 2Ce^\lambda$$

and so g is bounded as well. ►

2 The Stein–Chen method

Chen's characterisation of the Poisson suggests an approximation principle of some power and, in view of (1.6), we obtain a speculative refinement of the slogan with which we began.

SLOGAN *A random variable W is approximately Poisson if there is some λ for which $E(\lambda g(W + 1) - Wg(W)) \approx 0$ for bounded g .*

In detail, the procedure is as follows. Suppose $Z \sim \text{Poisson}(\lambda)$ where Z is defined on the same probability space as W . Suppose \mathbb{A} is any subset of the positive integers with $1_{\mathbb{A}}(k)$ its indicator. Write

$$P_\lambda(\mathbb{A}) = E 1_{\mathbb{A}}(Z) = P\{Z \in \mathbb{A}\} = \sum_{k \in \mathbb{A}} p(k; \lambda) \quad (2.1)$$

for the Poisson probability ascribed to the set \mathbb{A} . Now consider the indicator random variable $h(W) = 1_{\mathbb{A}}(W)$ governed by the probability law of W . If W is approximately Poisson(λ) (in a suitable sense) then we expect that $E(1_{\mathbb{A}}(W) - P_\lambda(\mathbb{A})) \approx 0$. By analogy with (1.6) and (1.7), this suggests that we seek a bounded $g(\cdot) = g_{\mathbb{A}}(\cdot)$ satisfying *Stein's equation*

$$\lambda g(k + 1) - kg(k) = 1_{\mathbb{A}}(k) - P_\lambda(\mathbb{A}) \quad (k \geq 0) \quad (2.2)$$

and anticipate that $E(\lambda g_{\mathbb{A}}(W + 1) - Wg_{\mathbb{A}}(W))$ should be small.

A natural notion of the goodness of fit between two distributions in this setting is the total variation distance. Suppose X and Y are arithmetic random variables. Write $\mathcal{L}(X)$ and $\mathcal{L}(Y)$, respectively, for their distributions (or *laws*).

DEFINITION The *total variation distance* between the laws of the arithmetic variables X and Y , denoted $d_{TV}(\mathcal{L}(X), \mathcal{L}(Y))$, is defined by

$$d_{TV}(\mathcal{L}(X), \mathcal{L}(Y)) = \sup_{\mathbb{A}} |P\{X \in \mathbb{A}\} - P\{Y \in \mathbb{A}\}|,$$

the supremum taken with respect to all subsets \mathbb{A} of the integers.

Specialising to our context, as $Z \sim \text{Poisson}(\lambda)$, we have $P\{Z \in \mathbb{A}\} = P_\lambda(\mathbb{A})$ and so

$$P\{W \in \mathbb{A}\} - P\{Z \in \mathbb{A}\} = E 1_{\mathbb{A}}(W) - P_\lambda(\mathbb{A}) = E(\lambda g_{\mathbb{A}}(W + 1) - Wg_{\mathbb{A}}(W))$$

in view of (2.2). The essence of the Stein–Chen method is hence contained in the following observation.

THEOREM *Suppose W is arithmetic and positive and Z is Poisson with mean λ . Then*

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) = \sup_{\mathbb{A}} |E(\lambda g_{\mathbb{A}}(W + 1) - Wg_{\mathbb{A}}(W))|$$

where $g_{\mathbb{A}}$ is a solution of Stein's equation (2.2).

To show that W is approximately Poisson it will suffice hence to obtain a small uniform bound for $E(\lambda g_{\mathbb{A}}(W+1) - Wg_{\mathbb{A}}(W))$, the size of the bound providing a measure of how well the law of W is approximated by a Poisson.

We now have a plausible, if indirect, framework for evaluating how well the slogan with which we began works in practice. The devil, of course, is in the details of the evaluation: the quality of the approximation depends upon the nature of the solution g of Stein's equation. Anticipating the need, we accordingly see what more can be gleaned from (2.2) before turning to a consideration of how the approximation may be prosecuted in practice.

3 Bounds from Stein's equation

Suppose \mathbb{A} is any subset of the positive integers and let $h(k) = 1_{\mathbb{A}}(k)$ be the indicator for \mathbb{A} . Also set $f(k) = 1_{\mathbb{A}}(k) - P_{\lambda}(\mathbb{A})$ where, in the notation of (2.1), $P_{\lambda}(\mathbb{A})$ is the Poisson set probability. We now set about specialising the explicit representation (1.8) to Stein's equation (2.2). Multiplying and dividing the right-hand side of (1.8) by e^{λ} we may write $g = g_{\mathbb{A}}$ in the form

$$g_{\mathbb{A}}(k+1) = \frac{e^{\lambda} k!}{\lambda^{k+1}} \sum_{j=0}^k p(j; \lambda) f(j) = \frac{1}{\lambda p(k; \lambda)} \sum_{j=0}^k p(j; \lambda) f(j) \quad (3.1)$$

where, as usual, $p(j; \lambda) = e^{-\lambda} \lambda^j / j!$ are the Poisson probabilities. Write $\mathbb{U}_k = \{0, 1, \dots, k\}$ for each k . Then, with $Z \sim \text{Poisson}(\lambda)$ and $f(Z) = 1_{\mathbb{A}}(Z) - P_{\lambda}(\mathbb{A})$, we may identify the sum on the right of (3.1) as the expectation of

$$1_{\mathbb{U}_k}(Z) f(Z) = 1_{\mathbb{A} \cap \mathbb{U}_k}(Z) - P_{\lambda}(\mathbb{A}) 1_{\mathbb{U}_k}(Z).$$

By taking expectations of both sides, we obtain

$$\begin{aligned} E(1_{\mathbb{U}_k}(Z) f(Z)) &= P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k) - P_{\lambda}(\mathbb{A}) P_{\lambda}(\mathbb{U}_k) \\ &= P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k) - P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k) P_{\lambda}(\mathbb{U}_k) - P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k^c) P_{\lambda}(\mathbb{U}_k) \\ &= P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k) P_{\lambda}(\mathbb{U}_k^c) - P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k^c) P_{\lambda}(\mathbb{U}_k) \end{aligned}$$

by two applications of additivity. It follows that we may write (3.1) in the form

$$g_{\mathbb{A}}(k+1) = \frac{1}{\lambda p(k; \lambda)} [P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k) P_{\lambda}(\mathbb{U}_k^c) - P_{\lambda}(\mathbb{A} \cap \mathbb{U}_k^c) P_{\lambda}(\mathbb{U}_k)] \quad (3.2)$$

from which we may deduce the useful fact that $g_{\mathbb{A}}(k) = -g_{\mathbb{A}^c}(k)$ for all \mathbb{A} . The relation (3.2) provides a splendid jump-off point for analysing the size and range of possible variation of the solutions g of Stein's equation. Reuse notation and write $\|g\| = \sup_k |g(k)|$.

THEOREM 1 Suppose g is any solution of (2.2). Then $\|g\| \leq 3 \min\{\lambda^{-1/2}, 1\}$.

PROOF: From (3.2) it is evident that

$$|g(k+1)| \leq \frac{\max\{P_\lambda(\mathbb{A} \cap \mathbb{U}_k)P_\lambda(\mathbb{U}_k^c), P_\lambda(\mathbb{A} \cap \mathbb{U}_k^c)P_\lambda(\mathbb{U}_k)\}}{\lambda p(k; \lambda)} \leq \frac{P_\lambda(\mathbb{U}_k)P_\lambda(\mathbb{U}_k^c)}{\lambda p(k; \lambda)}. \quad (3.3)$$

The inequality is actually tight as may be seen by setting $\mathbb{A} = \mathbb{U}_k$ or $\mathbb{A} = \mathbb{U}_k^c$.

Suppose first that $k \geq \lambda$. We may simplify the bound (3.3) further and write

$$\begin{aligned} |g(k+1)| &\leq \frac{P_\lambda(\mathbb{U}_k^c)}{\lambda p(k; \lambda)} = \frac{e^\lambda k! P\{Z \geq k+1\}}{\lambda^{k+1}} \\ &< \sqrt{\frac{2\pi}{k}} \left(\frac{k}{\lambda}\right)^{k+1} e^{\lambda-k+(12k)^{-1}} P\{Z \geq k+1\} \end{aligned} \quad (3.4)$$

in view of Robbins's upper bound for the factorial (XVI.6.1). Chernoff's bound for the Poisson tail helps wrap up. We have

$$\mathbf{E} e^{rZ} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j e^{rj}}{j!} = e^{-\lambda+\lambda e^r},$$

and so, for Chernoff's bound, we seek to optimise

$$\xi(r) = \mathbf{E} e^{r(Z-k-1)} = e^{-\lambda+\lambda e^r - r(k+1)}.$$

Differentiation shows that $f(r)$ is minimised when $r = \log(\frac{k+1}{\lambda}) \geq 0$ and so

$$\mathbf{P}\{Z \geq k+1\} \leq \inf_{r \geq 0} \mathbf{E} e^{r(Z-k-1)} = \xi\left(\log \frac{k+1}{\lambda}\right) = e^{-\lambda+k+1} \left(\frac{\lambda}{k+1}\right)^{k+1}.$$

Substitution in (3.4) then shows that

$$|g(k+1)| < \sqrt{\frac{2\pi}{k}} e^{1+(12k)^{-1}} \left(1 - \frac{1}{k+1}\right)^{k+1} \leq \sqrt{\frac{2\pi}{k}} e^{(12k)^{-1}}$$

in view of the basic identity $1-x \leq e^{-x}$. If $k \geq \lambda$ then k is at least one and so $\sqrt{2\pi} e^{(12k)^{-1}} \leq \sqrt{2\pi} e^{(12)^{-1}} < 3$ while $k^{-1/2} \leq \min\{1, \lambda^{-1/2}\}$ which shows that $|g(k+1)| < 3 \min\{\lambda^{-1/2}, 1\}$ for $k \geq \lambda$.

Suppose now that $k < \lambda$. If $k = 0$ then by direct calculation from (3.3), we have

$$|g(1)| \leq \frac{P_\lambda(\mathbb{U}_0)P_\lambda(\mathbb{U}_0^c)}{\lambda p(0; \lambda)} = \frac{1-e^{-\lambda}}{\lambda}.$$

The right-hand side is no larger than 1 (again, as $1-x \leq e^{-x}$) for all $\lambda > 0$; and if $\lambda \geq 1$ then it is certainly bounded above by $1/\lambda \leq 1/\sqrt{\lambda}$. If $\lambda < 1$ and $k < \lambda$

then $k = 0$ is the only case to consider and we are done. Suppose hence that $\lambda \geq 1$ and $1 \leq k < \lambda$. We now elect to bound (3.3) by

$$|g(k+1)| \leq \frac{P_\lambda(\mathbb{U}_k)}{\lambda p(k;\lambda)} = \frac{e^\lambda k! P\{Z \leq k\}}{\lambda^{k+1}} < \frac{\sqrt{2\pi k}}{\lambda} \left(\frac{k}{\lambda}\right)^k e^{\lambda-k+(12k)^{-1}} P\{Z \leq k\}.$$

Mimicking Chernoff's bound now for the left tail of the Poisson, we have

$$P\{Z \leq k\} \leq \inf_{r \geq 0} E e^{-r(Z-k)} = e^{-\lambda+k} \left(\frac{\lambda}{k}\right)^k.$$

When $1 \leq k < \lambda$ we hence obtain the bound

$$|g(k+1)| < \frac{\sqrt{2\pi k}}{\lambda} e^{(12k)^{-1}} < \sqrt{\frac{2\pi}{\lambda}} e^{(12)^{-1}} < 3 \min\{\lambda^{-1/2}, 1\}.$$

The two cases considered cover all possibilities for k , concluding the proof. ▶

Of even more interest than the maximum size of g is the rate of change given by the maximum of the successive differences, $\Delta g = \sup_k |g(k+1) - g(k)|$.

THEOREM 2 *If g is any solution of (2.2) then $\Delta g \leq \lambda^{-1}(1 - e^{-1}) \leq \min\{\lambda^{-1}, 1\}$.*

PROOF: Write $f(k) = 1_{\mathbb{A}} - P_\lambda(\mathbb{A}) = \sum_{j \in \mathbb{A}} (1_{\{j\}}(k) - P_\lambda\{j\})$. Here, of course, $P_\lambda\{j\} = p(j;\lambda)$ are just the Poisson point probabilities. By additivity, we may hence decompose the solution of Stein's equation (2.2) in the form

$$g_{\mathbb{A}}(k) = \sum_{j \in \mathbb{A}} g_{\{j\}}(k). \quad (3.5)$$

We are hence led to consider the point solutions $g_{\{j\}}(k)$ of Stein's equation. Fix any positive integer j and identify $\mathbb{A} = \{j\}$ as a singleton in (3.2). We then have

$$g_{\{j\}}(k+1) = \begin{cases} \frac{-p(j;\lambda) P\{Z \leq k\}}{\lambda p(k;\lambda)} & \text{if } 0 \leq k \leq j-1, \\ \frac{p(j;\lambda) P\{Z \geq k+1\}}{\lambda p(k;\lambda)} & \text{if } k \geq j. \end{cases} \quad (3.6)$$

Writing down the point probabilities explicitly, we see that

$$\frac{P\{Z \leq k\}}{p(k;\lambda)} = \sum_{m=0}^k \frac{\lambda^{-(k-m)} k!}{m!} \stackrel{(n=k-m)}{=} \sum_{n=0}^k \frac{\lambda^{-n} k!}{(k-n)!} = \sum_{n=0}^k \lambda^{-n} k^n$$

in the falling factorial notation $k^n = k(k-1)\cdots(k-n+1)$. Both the number of terms in the sum and the summands themselves increase with k and so $P\{Z \leq k\}/p(k;\lambda)$ increases monotonically as k increases. By a similar exercise,

$$\frac{P\{Z \geq k+1\}}{p(k;\lambda)} = \sum_{m=k+1}^{\infty} \frac{\lambda^{m-k} k!}{m!} \stackrel{(n=m-k)}{=} \sum_{n=1}^{\infty} \frac{\lambda^n k!}{(k+n)!} = \sum_{n=1}^{\infty} \frac{\lambda^n}{(k+n)^n}$$

and, as the terms of the sum decrease with k , it follows that $P\{Z \geq k+1\}/p(k; \lambda)$ decreases monotonically as k increases. Thus, $g_{\{j\}}(k+1)$ is negative and decreasing as k increases from 0 to $j-1$, and is positive and decreasing as k increases beyond j . It follows that, if $j > 0$, the successive differences $g_{\{j\}}(k+1) - g_{\{j\}}(k)$ are negative *excepting only* when $k = j$. Focusing hence on the point $k = j$, the representation (3.6) yields

$$\begin{aligned} 0 \leq g_{\{j\}}(j+1) - g_{\{j\}}(j) &= \frac{p(j; \lambda) P\{Z \geq j+1\}}{\lambda p(j; \lambda)} + \frac{p(j; \lambda) P\{Z \leq j-1\}}{\lambda p(j-1; \lambda)} \\ &= \frac{P\{Z \geq j+1\}}{\lambda} + \frac{P\{Z \leq j-1\}}{j} = \frac{e^{-\lambda}}{\lambda} \left[\sum_{i=j+1}^{\infty} \frac{\lambda^i}{i!} + \sum_{i=0}^{j-1} \frac{\lambda^{i+1}}{(i+1)!} \cdot \frac{i+1}{j} \right] \\ &\leq \frac{e^{-\lambda}}{\lambda} \min \left\{ \sum_{i=1}^{\infty} \frac{\lambda^i}{i!}, \sum_{i=1}^{\infty} \frac{\lambda^i}{i!} \cdot \frac{i}{j} \right\} = \min \{ \lambda^{-1}(1 - e^{-\lambda}), j^{-1} \} \leq \lambda^{-1}(1 - e^{-\lambda}) \end{aligned}$$

as $1 - e^{-\lambda} \leq \lambda$ and $j \geq 1$. For $j = 0$ the successive differences $g_{\{0\}}(k+1) - g_{\{0\}}(k)$ are all negative. In view of the decomposition (3.5), we hence see that

$$g_{\mathbb{A}}(k+1) - g_{\mathbb{A}}(k) = \sum_{j \in \mathbb{A}} (g_{\{j\}}(k+1) - g_{\{j\}}(k)) \leq \lambda^{-1}(1 - e^{-\lambda})$$

as at most one of the summands is positive as j runs through \mathbb{A} . As $g_{\mathbb{A}} = -g_{\mathbb{A}^c}$, by applying the just-derived bound to the set \mathbb{A}^c , we also have

$$g_{\mathbb{A}}(k+1) - g_{\mathbb{A}}(k) = -(g_{\mathbb{A}^c}(k+1) - g_{\mathbb{A}^c}(k)) \geq -\lambda^{-1}(1 - e^{-\lambda}).$$

By pooling the bounds on the successive differences, we obtain

$$\Delta g_{\mathbb{A}} = \sup_k |g_{\mathbb{A}}(k+1) - g_{\mathbb{A}}(k)| \leq \lambda^{-1}(1 - e^{-\lambda}) \leq \min\{\lambda^{-1}, 1\},$$

the bounds uniform in \mathbb{A} as was to be shown. ▶

The appearance of the magic factor λ^{-1} bounding Δg is the key to obtaining good approximations.

4 Sums of indicators

The natural domain of Poisson approximation is in the analysis of extremes and exceedances. We consider the setting of a finite family of indicator variables $\{X_{\alpha}, \alpha \in \Gamma\}$ where, for each α , the variable X_{α} has mean p_{α} . The finite index set Γ is usually taken to be $\{1, \dots, n\}$ though on occasion it will be useful to notationally permit other enumerations than the ordinal. In typical settings the indicator variables X_{α} track the occurrence of rare or extreme events. The sum $W = \sum_{\alpha} X_{\alpha}$ then indicates the number of such occurrences.

The intuition built up from the methods of Chapter IV suggests that if the dependence is not too severe between the events for which X_α are the indicators then W should be governed approximately by a Poisson distribution. It may, however, be quite difficult to capture the dependency structure sufficiently accurately for sieve methods to be effective; the reader will recall that calculations of some subtlety are required even in the relatively simple situations seen in Chapter IV.

The programme outlined in Section 2 on the other hand depends for its success on the straightforward identity (1.6). The procedure, however, is certainly not intuitive nor is it clear whether it provides any tangible gains. We should certainly improve our understanding of the process by starting with some simple examples. We begin accordingly with the familiar setting of independent trials.

Suppose $\{X_\alpha, \alpha \in \Gamma\}$ is a family of *independent* indicator variables. The distribution of $W = \sum_\alpha X_\alpha$ is then given by repeated convolutions of Bernoulli distributions but it is too much to hope for a closed-form solution in the case of non-identical trials.

For each α , it will be convenient to introduce the nonce notation $W_{(\alpha)} = \sum_{\beta \neq \alpha} X_\beta = W - X_\alpha$ for the accumulated number of successes with trial α expurgated. Then, for any bounded g , by additivity of expectation, we have

$$\mathbb{E}(Wg(W)) = \sum_\alpha \mathbb{E}(X_\alpha g(W_{(\alpha)}) + X_\alpha).$$

Now, by conditioning on the result of trial α , we obtain

$$\mathbb{E}(X_\alpha g(W_{(\alpha)} + X_\alpha)) = p_\alpha \mathbb{E}(g(W_{(\alpha)} + 1) | X_\alpha = 1) \stackrel{(*)}{=} p_\alpha \mathbb{E}(g(W_{(\alpha)} + 1)), \quad (4.1)$$

the step marked $(*)$ justified by the fact that $W_{(\alpha)}$ and X_α are independent. It follows that

$$\mathbb{E}(Wg(W)) = \sum_\alpha p_\alpha \mathbb{E}(g(W_{(\alpha)} + 1)).$$

If Poisson convergence is on the cards at all we should set λ to be the mean of W . Accordingly, we set $\lambda = \mathbb{E}(W) = \sum_\alpha p_\alpha$ whence, by additivity, we have

$$\mathbb{E}(\lambda g(W + 1)) = \sum_\alpha p_\alpha \mathbb{E} g(W_{(\alpha)} + X_\alpha + 1).$$

Collecting terms and simplifying further by conditioning once more on the results of the individual trials yields

$$\begin{aligned} \mathbb{E}(\lambda g(W + 1) - Wg(W)) &= \sum_\alpha p_\alpha \mathbb{E}(g(W_{(\alpha)} + X_\alpha + 1) - g(W_{(\alpha)} + 1)) \\ &= \sum_\alpha p_\alpha^2 \mathbb{E}(g(W_{(\alpha)} + 2) - g(W_{(\alpha)} + 1) | X_\alpha = 1) = \sum_\alpha p_\alpha^2 \mathbb{E}(g(W_{(\alpha)} + 2) - g(W_{(\alpha)} + 1)) \end{aligned}$$

where we again utilise the fact that $W_{(\alpha)}$ is independent of X_α . Specialise now to any solution g of Stein's equation. By taking absolute values we then have

$$|\mathbb{E}(\lambda g(W + 1) - Wg(W))| \leq \Delta g \sum_\alpha p_\alpha^2$$

which, in view of Theorem 3.2, provides explicit bounds on the quality of Poisson approximation.

THEOREM Suppose $\{X_\alpha, \alpha \in \Gamma\}$ is a finite sequence of (independent) Bernoulli trials where, for each $\alpha \in \Gamma$, X_α has mean p_α . Let $W = \sum_\alpha X_\alpha$ and suppose Z is a Poisson variable with mean $\lambda = \sum_\alpha p_\alpha$. Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \lambda^{-1}(1 - e^{-\lambda}) \sum_\alpha p_\alpha^2 \leq \min \left\{ \frac{\sum_\alpha p_\alpha^2}{\sum_\alpha p_\alpha}, \sum_\alpha p_\alpha^2 \right\}.$$

In the case of identical Bernoulli trials the situation reverts to the familiar. Suppose $\Gamma = \{1, \dots, n\}$ and $p_1 = \dots = p_n = r$. Then W has the binomial distribution $b_n(k; r) = \binom{n}{k} r^k (1-r)^{n-k}$ and the theorem addresses the accuracy of the Poisson approximation to the binomial.

COROLLARY Suppose $n \geq 1$ and $0 < r < 1$. Then the total variation distance between the Binomial(n, r) and Poisson(nr) distributions is no larger than $\min\{r, nr^2\}$. It follows a fortiori that $|b_n(k; r) - p(k; nr)| \leq \min\{r, nr^2\}$ for every k .

If $r = r_n \rightarrow 0$ then the Poisson distribution gives a uniformly good approximation to the binomial. This is already much stronger than the classical Poisson approximation of Theorem IV.6.3 which only asserts pointwise (but not uniform) convergence of point probabilities under the condition that the sequence $\{nr_n, n \geq 1\}$ converges. The strengthening of Poisson convergence in the corollary not only expands widely the range of applicability of Poisson's theorem but provides explicit error bounds. This is encouraging; the Stein–Chen method has effortlessly improved upon the classical result. The theorem moreover provides cover for Poisson approximation for sums of independent indicators with variable success probabilities, a setting for which we had no prior results at all. But the largest benefits are still to appear—the setting extends seamlessly to situations where there is dependence across the indicator events.

Suppose now that the indicator random variables X_α have some dependency structure. We should then try to retrace the previous steps without the independence assumption. If the reader attempts this she will find that the step marked $(*)$ in (4.1) breaks down. A little thought will suggest a way out: the analysis remains impeccable if we simply retain the conditioning on the results of the individual trials. By additivity of expectation we may write

$$\begin{aligned} E(\lambda g(W+1) - Wg(W)) &\stackrel{(i)}{=} \sum_\alpha [p_\alpha E(g(W+1)) - E(X_\alpha g(W))] \\ &\stackrel{(ii)}{=} \sum_\alpha p_\alpha [E(g(W+1)) - E(g(W_{(\alpha)} + 1) | X_\alpha = 1)]. \end{aligned} \quad (4.2)$$

This formulation is promising in two generic settings: (i) dependencies across the X_α , while possibly strong, are sparse; and (ii) the variables X_α have a pervasive dependency structure but the dependence is weak. We consider these in turn.

5 The local method, dependency graphs

The simplest approach to computing the sums on the right in (4.2) is to attempt to leverage independence whenever possible so that the previous analysis is applicable to these terms and attempt to bound the remaining dependent cases in the sum. If we are fortunate enough that the dependent terms in the sum are dominated by the independent terms then the analysis for the case of independent trials will more or less go through. This method was originally exploited by Chen in 1975.

The procedure is as follows. For each $\alpha \in \Gamma$, we first identify a subset of vertices Γ'_α ; these are the indices β of those variables X_β that are “strongly” dependent on X_α . The set $\Gamma_\alpha = \Gamma \setminus (\Gamma'_\alpha \cup \{\alpha\})$ then identifies the indices of variables X_β which are only “weakly” dependent on X_α . In typical settings we want $\text{card } \Gamma'_\alpha$ to be small and X_α to be nearly independent of the collection $\{X_\beta, \beta \in \Gamma_\alpha\}$.

For each α , let $T_\alpha = \sum_{\beta \in \Gamma'_\alpha} X_\beta$ and $S_\alpha = \sum_{\beta \in \Gamma_\alpha} X_\beta$. The idea now is to isolate the contribution from strongly dependent terms. As $W = S_\alpha + T_\alpha + X_\alpha$, by telescoping terms, we have

$$\begin{aligned} E(X_\alpha g(W)) &= E(X_\alpha g(S_\alpha + T_\alpha + X_\alpha)) = E(X_\alpha g(S_\alpha + T_\alpha + 1)) \\ &= p_\alpha E(g(S_\alpha + 1)) + E((X_\alpha - p_\alpha)g(S_\alpha + 1)) + E(X_\alpha[g(S_\alpha + T_\alpha + 1) - g(S_\alpha + 1)]). \end{aligned}$$

Substituting into the step marked (i) in (4.2), we may decompose the sum into

$$\begin{aligned} E(\lambda g(W+1) - Wg(W)) &= \sum_\alpha \underbrace{\left[p_\alpha E(g(X_\alpha + S_\alpha + T_\alpha + 1) - g(S_\alpha + 1)) \right]}_A \\ &\quad - \underbrace{E((X_\alpha - p_\alpha)g(S_\alpha + 1))}_B - \underbrace{E(X_\alpha[g(S_\alpha + T_\alpha + 1) - g(S_\alpha + 1)])}_C. \end{aligned}$$

The terms A and C are set up nicely for estimation in terms of the successive differences of g . Begin with the observation

$$g(m+k) - g(m) = \sum_{j=0}^{k-1} [g(m+j+1) - g(m+j)]$$

as the series telescopes. It follows hence by the triangle inequality that

$$|g(m+k) - g(m)| \leq \sum_{j=0}^{k-1} |g(m+j+1) - g(m+j)| \leq k \Delta g \quad (5.1)$$

for every choice of m and k . We hence obtain the simple bounds

$$|A| \leq \Delta g \cdot p_\alpha E(X_\alpha + T_\alpha) = \Delta g \cdot p_\alpha \left(p_\alpha + \sum_{\beta \in \Gamma'_\alpha} p_\beta \right),$$

$$|C| \leq \Delta g \cdot E(X_\alpha T_\alpha) = \Delta g \sum_{\beta \in \Gamma'_\alpha} E(X_\alpha X_\beta).$$

The term B is the spoiler consisting of the “nearly independent” contributions. As the collection $\{X_\beta, \beta \in \Gamma_\alpha\}$ completely determines $S_\alpha = \sum_{\beta \in \Gamma_\alpha} X_\beta$, by conditioning on the values of the indicators X_β for $\beta \in \Gamma_\alpha$, we have

$$E((X_\alpha - p_\alpha)g(S_\alpha + 1) | X_\beta, \beta \in \Gamma_\alpha) = E((X_\alpha - p_\alpha) | X_\beta, \beta \in \Gamma_\alpha)g(S_\alpha + 1).$$

Removing the conditioning by taking expectation with respect to $\{X_\beta, \beta \in \Gamma_\alpha\}$ we see that

$$B = E[(E(X_\alpha | X_\beta, \beta \in \Gamma_\alpha) - p_\alpha)g(S_\alpha + 1)],$$

so that, in view of the modulus inequality $|E(X)| \leq E|X|$, we have

$$|B| \leq \|g\| \cdot E|E(X_\alpha | X_\beta, \beta \in \Gamma_\alpha) - p_\alpha|.$$

The bounds on A, B, and C are simplified by appeal to Theorems 3.1, 2 to bound the variation of g. Compacting expressions by identifying

$$\Sigma_1 = \sum_\alpha E(X_\alpha)^2, \quad \Sigma_2 = \sum_\alpha \sum_{\beta \in \Gamma'_\alpha} E(X_\alpha) E(X_\beta), \quad \Sigma_3 = \sum_\alpha \sum_{\beta \in \Gamma'_\alpha} E(X_\alpha X_\beta), \quad (5.2)$$

$$\Sigma_4 = \sum_\alpha E|E(X_\alpha | X_\beta, \beta \in \Gamma_\alpha) - E(X_\alpha)| \quad (5.3)$$

helps keep focus on the key rates of growth and we finally obtain

$$|E(\lambda g(W+1) - Wg(W))| \leq \lambda^{-1}(1 - e^{-\lambda})(\Sigma_1 + \Sigma_2 + \Sigma_3) + 3 \min\{\lambda^{-1/2}, 1\} \cdot \Sigma_4.$$

A bound on the quality of Poisson approximation follows immediately. As before, let Z be a Poisson variable with mean $\lambda = \sum_\alpha p_\alpha = \sum_\alpha E(X_\alpha)$. I will opt for slightly less precision in favour of a compact presentation.

THEOREM Suppose $\{\Gamma_\alpha, \Gamma'_\alpha\}$ is a partition of $\Gamma \setminus \{\alpha\}$ for each α . Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \lambda^{-1}(\Sigma_1 + \Sigma_2 + \Sigma_3) + 3\lambda^{-1/2}\Sigma_4. \quad (5.4)$$

The terms Σ_2 , Σ_3 and, particularly, Σ_4 are the “dependence penalty” we pay over the theorem of the previous section. What makes this bound promising is that Σ_1 , Σ_2 , and Σ_3 depend only on the first two moments of the indicator variables. The contribution from Σ_4 is more involved and gives poorer approximations because it is damped only by $\lambda^{-1/2}$ instead of λ^{-1} . But if we can choose the partitions such that, for each α , the indicator X_α is almost independent of the family $\{X_\beta, \beta \in \Gamma_\alpha\}$ then we can force this term to be small. The simplest such setting is when, in the terminology of Section IV.9, Γ_α is the independence set associated with a dependency graph on the vertex set Γ ; in this case there is a purely local dependency structure captured by the adjacency sets Γ'_α .

In detail, we construct a directed dependency graph \mathcal{G} on the set of vertices Γ as follows. As in Section IV.9, the edges of \mathcal{G} are determined by the local dependency structure of the variables X_α . For each α , let Γ_α denote a maximal set of vertices β in Γ for which X_α is independent of the subfamily $\{X_\beta, \beta \in \Gamma_\alpha\}$; this is the *independence set* associated with α . Let $\Gamma'_\alpha = \Gamma \setminus (\Gamma_\alpha \cup \{\alpha\})$. Then $\{\Gamma_\alpha, \Gamma'_\alpha\}$ forms a partition of $\Gamma \setminus \{\alpha\}$. To each vertex α in \mathcal{G} we now associate the incident edges (α, β) with $\beta \in \Gamma'_\alpha$. The local structure of the graph \mathcal{G} in a natural sense identifies the local dependency structure of the variables X_α . By construction, X_α is independent of the subfamily $\{X_\beta, \beta \in \Gamma_\alpha\}$ and so $\Sigma_4 = 0$.

The local method is most powerful when the dependency structure can be clearly identified to be sparse. Very roughly speaking, Poisson approximation via the local method will yield good results if dependency is very local and the accumulated effect of the correlations is hence small. An old acquaintance provides an illustrative setting.

6 Triangles and cliques in random graphs, reprise

Consider the Erdős–Rényi random graph $G_{n,p}$. Modifying the notation of Section XVI.4 to fit the current context, let α denote a generic subgraph on three vertices, Γ the family of such subgraphs. Let X_α be the indicator for the event that α forms a triangle. Then $W = \sum_\alpha X_\alpha$ denotes the number of triangles in $G_{n,p}$. If $p = p_n$ is small we anticipate that W will be approximately Poisson with mean np_n .

For each α , the independence set Γ_α in the induced dependency graph consists of those subgraphs β on three vertices which share one or fewer vertices with α , the indicator X_α independent of the family $\{X_\beta, \beta \in \Gamma_\alpha\}$. The local dependency structure is hence captured by the set $\Gamma'_\alpha = \Gamma_\alpha \setminus (\Gamma_\alpha \cup \{\alpha\})$; these are the subgraphs β on three vertices that share precisely two vertices with α . Thus, the induced dependency graph has degree $\text{card } \Gamma'_\alpha = \binom{3}{2} \binom{n-3}{1} = 3(n-3)$. As a subgraph α on three vertices of $G_{n,p}$ forms a triangle if, and only if, all three edges are present, $E(X_\alpha) = p^3$ for each α . Likewise, if $\beta \in \Gamma'_\alpha$, the subgraphs α and β each form triangles if, and only if, the five distinct edges in play are all present (see Figure XVI.4.1) and so $E(X_\alpha X_\beta) = p^5$. Running through the calculations in (5.2), we hence obtain

$$\begin{aligned}\lambda &= \sum_\alpha E(X_\alpha) = \binom{n}{3} p^3 = \frac{n^3 p^3}{6}, & \Sigma_1 &= \sum_\alpha E(X_\alpha)^2 = \binom{n}{3} p^6 = \frac{n^3 p^6}{6}, \\ \Sigma_2 &= \sum_\alpha \sum_{\beta \in \Gamma'_\alpha} E(X_\alpha) E(X_\beta) = \binom{n}{3} \binom{3}{2} \binom{n-3}{1} p^6 = \frac{n^4 p^6}{2}, \\ \Sigma_3 &= \sum_\alpha \sum_{\beta \in \Gamma'_\alpha} E(X_\alpha X_\beta) = \binom{n}{3} \binom{3}{2} \binom{n-3}{1} p^5 = \frac{n^4 p^5}{2},\end{aligned}$$

the falling factorial notation $n^k = n(n-1)\cdots(n-k+1)$ compacting expressions. As $n^k \sim n^k$ for each fixed k , the term $\Sigma_3 \sim n^4 p^5 / 2$ dominates $\Sigma_1 + \Sigma_2$ when $p = p_n$ is small. As X_α is independent of the family $\{X_\beta, \beta \in \Gamma_\alpha\}$ by substitution in (5.3) it follows that $\Sigma_4 = 0$. Thus, when $p = p_n$ is small, the dominant term on the right in (5.4) is $\lambda^{-1} \Sigma_3$ and we obtain $d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) = \mathcal{O}(np^2)$ eventually (we shall not bother being too careful with the constants).

THEOREM 1 *Let W be the number of triangles in $G_{n,p}$ and let Z be Poisson with mean $\lambda = \binom{n}{k} p^{k(k-1)/2}$. If $p = p_n = o(n^{-1/2})$, then $d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \rightarrow 0$ as $n \rightarrow \infty$ and, a fortiori, if $p = p_n = c/n$ for any positive constant c then the number of triangles in G_{n,p_n} is governed asymptotically by a Poisson distribution of mean $c^3/6$.*

In particular, the attribute that $G_{n,p}$ has a triangle exhibits a phase transition as p increases past the order of $1/n$. This effortless strengthening of the theorem of Section XVI.4 has needed only second moments!



EXTENSION: CLIQUES

If the reader considers the argument she will realise that it is clear that we could replace the attribute that “ $G_{n,p}$ has a triangle” by “ $G_{n,p}$ has a clique on k vertices” (that is to say, a complete subgraph on k vertices) without changing the essential nature of the analysis. All that is required are second moment calculations, slightly messier to be sure but posing no insuperable barrier, and again these are dominated by the mean. I will sketch the argument.

Let Γ now denote the family of subgraphs on k vertices of $G_{n,p}$. Then $\text{card } \Gamma = \binom{n}{k} = \frac{n^k}{k!} \sim \frac{n^k}{k!}$. For $\alpha \in \Gamma$ let X_α be the indicator for the event that α forms a complete subgraph on k vertices. As there are $\binom{k}{2}$ potential edges in play between the k vertices in α , we have $E(X_\alpha) = p^{\binom{k}{2}}$. Accordingly,

$$\lambda = \sum_{\alpha} E(X_\alpha) = \binom{n}{k} p^{\binom{k}{2}} \sim \frac{n^k p^{k(k-1)/2}}{k!}.$$

The dependency structure is sparse and so we anticipate a threshold function for the emergence of a clique when $p = p_n$ is of the asymptotic order of $n^{-2/(k-1)}$.

If α and β share fewer than two vertices then the indicators X_α and X_β are independent as α and β will share no edges. Accordingly, to set up the local method let Γ_α now denote the independence set of subgraphs β on k vertices that share 0 or 1 vertices with α . The bookend sums in (5.2,5.3) are straightforward and we dispose of them first. To begin,

$$\Sigma_1 = \sum_{\alpha} E(X_\alpha)^2 = \binom{n}{k} p^{2\binom{k}{2}} \sim \frac{n^k p^{k(k-1)}}{k!},$$

and as X_α is independent of the family $\{X_\beta, \beta \in \Gamma_\alpha\}$ we have $\Sigma_4 = 0$. Now to estimate the remaining sums.

We may partition $\Gamma'_\alpha = \Gamma \setminus (\Gamma_\alpha \cup \{\alpha\})$ into disjoint sets $\Gamma'_\alpha(2), \Gamma'_\alpha(3), \dots, \Gamma'_\alpha(k-1)$ where, for $2 \leq \ell \leq k-1$, $\Gamma'_\alpha(\ell)$ denotes the family of subgraphs β on k vertices which

share ℓ vertices with α . As the shared vertices can be specified in $\binom{k}{\ell}$ ways and the remaining vertices in $\binom{n-k}{k-\ell}$ ways, we have $\text{card } \Gamma'_\alpha(\ell) = \binom{k}{\ell} \binom{n-k}{k-\ell}$. For fixed k and ℓ ,

$$\binom{n-k}{k-\ell} = \frac{(n-k)(n-k-1)\cdots(n-2k+\ell+1)}{(k-\ell)!} \sim \frac{n^{k-\ell}}{(k-\ell)!}$$

asymptotically in n as, by factoring out n from each of the $k-\ell$ terms in the numerator what remains is a product of $k-\ell$ terms of the form $1-j/n$ each of which is close to 1. Thus, $\text{card } \Gamma'_\alpha(\ell) \sim \binom{k}{\ell} \frac{n^{k-\ell}}{(k-\ell)!}$. We now have

$$\begin{aligned} \Sigma_2 &= \sum_{\alpha} \sum_{\beta \in \Gamma'_\alpha} \mathbb{E}(X_\alpha) \mathbb{E}(X_\beta) = \sum_{\alpha} \sum_{2 \leq \ell \leq k-1} \sum_{\beta \in \Gamma'_\alpha(\ell)} \mathbb{E}(X_\alpha) \mathbb{E}(X_\beta) \\ &= \binom{n}{k} \sum_{\ell=2}^{k-1} \text{card } \Gamma'_\alpha(\ell) p^{2\binom{k}{2}} = \sum_{\ell=2}^{k-1} \binom{n}{k} \binom{k}{\ell} \binom{n-k}{k-\ell} p^{2\binom{k}{2}} \sim \frac{n^{2(k-1)} p^{k(k-1)}}{2(k-2)!^2} \end{aligned}$$

as the dominant term in the sum corresponds to $\ell = 2$ when the exponent of n is largest. Now, if $\beta \in \Gamma'_\alpha(\ell)$ then the number of edges in play between α and β is $2\binom{k}{2} - \binom{\ell}{2}$ as the edges connecting the ℓ vertices in common appear in both α and β . Thus, $\mathbb{E}(X_\alpha X_\beta) = p^{2\binom{k}{2} - \binom{\ell}{2}}$. Repeating the calculations for Σ_2 with the product of means $\mathbb{E}(X_\alpha) \mathbb{E}(X_\beta)$ in the summands replaced by the correlations $\mathbb{E}(X_\alpha X_\beta)$, we obtain

$$\Sigma_3 = \sum_{\ell=2}^{k-1} \binom{n}{k} \binom{k}{\ell} \binom{n-k}{k-\ell} p^{2\binom{k}{2} - \binom{\ell}{2}} \sim \sum_{\ell=2}^{k-1} \frac{\binom{k}{\ell}}{k!(k-\ell)!} n^{2k-\ell} p^{k(k-1)-\ell(\ell-1)/2}. \quad (6.1)$$

The summand for $\ell = 2$ by itself dominates Σ_1 and Σ_2 and so Σ_3 is dominant in the sum in (5.4). In order to determine the asymptotic behaviour of Σ_3 observe that as ℓ varies the behaviour of the summands on the right in (6.1) is determined by the term $n^{-\ell} p^{-\ell(\ell-1)/2}$. Accordingly, consider the function

$$f(x) = n^{-x} p^{-x(x-1)/2} = \exp(-x \log n - \frac{x(x-1)}{2} \log p).$$

Differentiation shows that f has a unique minimum at $x = \frac{1}{2}(1 - \frac{2 \log n}{\log p})$. It follows that the dominant term in the sum on the right in (6.1) occurs at one end or the other, $\ell = 2$ or $\ell = k-1$, of the sum. Accordingly, with a fine disregard for constants, we write

$$\Sigma_3 = \mathcal{O}(n^{2k-2} p^{k(k-1)-2(2-1)/2} + n^{2k-(k-1)} p^{k(k-1)-(k-1)(k-2)/2}).$$

By pooling estimates we hence obtain

$$\lambda^{-1}(\Sigma_1 + \Sigma_2 + \Sigma_3) + 3\lambda^{-1/2}\Sigma_4 \sim \lambda^{-1}\Sigma_3 = \mathcal{O}(n^{k-2} p^{(k-2)(k+1)/2} + np^{k-1}).$$

If we set $p = p_n = cn^{-2/(k-1)}$ for some positive constant c then

$$n^{k-2} p^{(k-2)(k+1)/2} + np^{k-1} = c_1 n^{-2(k-2)/(k-1)} + c_2 n^{-1} \leq (c_1 + c_2) n^{-1}$$

as $(k-2)/(k-1) = 1 - 1/(k-1) \geq 1/2$ for $k \geq 3$. Here $c_1 = c^{(k-2)(k+1)/2}$ and $c_2 = c^{k-1}$ are constants determined by c and k but their explicit form is not germane for our purposes; what matters is the asymptotic rate of decay which leads to the following

THEOREM 2 Suppose $p = p_n = cn^{-2/(k-1)}$ for some positive constant c . Let W be the number of cliques of size k in $G_{n,p}$ and let Z be a Poisson variable with mean $c^{k(k-1)/2}/k!$. Then $d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) = \Theta(n^{-1})$ and, a fortiori, the point probabilities for W converge pointwise to that of the Poisson with mean $c^{k(k-1)/2}/k!$.

Thus, if $p = p_n = cn^{-2/(k-1)}$, the probability that there are no cliques on k vertices in $G_{n,p}$ tends asymptotically to $\exp(-c^{k(k-1)/2}/k!)$. If the positive c is small then $P\{W=0\} \approx 1$; if, on the other hand, c is large then $P\{W=0\} \approx 0$. There is hence a phase transition in the appearance of cliques on k vertices as $p = p_n$ increases past the order of $n^{-2/(k-1)}$. This generalises our earlier result from triangles to cliques.

7 Pervasive dependence, the method of coupling

Chen's original approach to exploiting local structure is less likely to be successful when dependencies are pervasive but weak.

EXAMPLE: *Revisiting le problème des rencontres.* Let (Π_1, \dots, Π_n) be a random permutation of $(1, \dots, n)$. The index set here is $\Gamma = \{1, \dots, n\}$. For each $\alpha \in \Gamma$, let $X_\alpha = 1(\Pi_\alpha = \alpha)$ be the indicator for the event that the position of the element α is unchanged in the permutation, and let $W = \sum_\alpha X_\alpha$ be the number of matches. This is de Montmort's problème des rencontres that we analysed by sieve methods in Example IV.6.3.

By the symmetry of the problem it is clear that the random variables (X_1, \dots, X_n) are *exchangeable*, that is to say, their joint distribution $F(x_1, \dots, x_n)$ is invariant with respect to any permutation of the arguments. In particular, we have $E(X_\alpha) = 1/n$ for all α and $E(X_\alpha X_\beta) = 1/n(n-1)$ if $\alpha \neq \beta$. For each α , the independence set Γ_α of the dependency graph is hence empty and $\Gamma'_\alpha = \Gamma \setminus \{\alpha\}$ has cardinality $n-1$. Substituting in (5.2) and (5.3), it follows that

$$\begin{aligned}\lambda &= n \cdot \frac{1}{n} = 1, & \Sigma_1 &= n \cdot \frac{1}{n^2} = \frac{1}{n}, & \Sigma_2 &= n(n-1) \cdot \frac{1}{n^2} = 1 - \frac{1}{n}, \\ \Sigma_3 &= n(n-1) \cdot \frac{1}{n(n-1)} = 1, & \text{and} & & \Sigma_4 &= 0.\end{aligned}$$

Plugging into (5.4) we obtain $d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq 2$ and the bound is useless. ▶

This type of setting is better suited to a rather different approach proposed by Stein in his monograph. This approach goes under the umbrella identification “the method of coupling” which describes a vague but powerful idea of replacing a complex calculation by an equivalent but simpler structure. I shall extract the key elements of the method and present characteristic applications indicative of its utility in this and the following sections. My presentation follows the excellent monograph of Barbour, Holst, and Janson which fleshes out the method and presents its range of application in a clear light.⁷

⁷A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation*. Oxford: Oxford University Press, 1992.

To set the stage, suppose Γ is a finite index set and $\{X_\alpha, \alpha \in \Gamma\}$ is a family of (possibly dependent) indicator variables where, for each α , X_α has mean p_α . As before, we write $W = \sum_\alpha X_\alpha$ and set $\lambda = E(W) = \sum_\alpha p_\alpha$. We suppose that Z is a Poisson random variable with mean λ .

Suppose now that we select random variables U_α and V_α (on the same probability space) so that, for each α , U_α has the same distribution as W (a natural choice in many situations is $U_\alpha = W$ though it is useful to retain flexibility) and $V_\alpha + 1$ has the distribution of W conditioned on $X_\alpha = 1$ (or, equivalently, V_α has the distribution of $W_{(\alpha)}$ conditioned on $X_\alpha = 1$). We indicate this in notation by writing

$$\mathcal{L}(U_\alpha) = \mathcal{L}(W) \text{ and } \mathcal{L}(V_\alpha) = \mathcal{L}(W_{(\alpha)} | X_\alpha = 1). \quad (7.1)$$

The random variables U_α and V_α are the *couplings*. As we shall see in the applications, a proper choice of couplings can dramatically simplify the estimation.

For any choice of couplings with the appropriate distributions, by substituting into the step marked (ii) in (4.2), we obtain

$$E(\lambda g(W+1) - Wg(W)) = \sum_\alpha p_\alpha E(g(U_\alpha + 1) - g(V_\alpha + 1)). \quad (7.2)$$

We may simplify the right-hand side by the observation (5.1). As

$$|g(U_\alpha + 1) - g(V_\alpha + 1)| \leq |U_\alpha - V_\alpha| \Delta g,$$

taking absolute values of both sides of (7.2) leads to

$$|E(\lambda g(W+1) - Wg(W))| \leq \Delta g \sum_\alpha p_\alpha E|U_\alpha - V_\alpha|.$$

Specialising to solutions of Stein's equation, Theorem 3.2 yields a coupling bound on the quality of Poisson approximation.

THEOREM 1 Suppose that, for each α , U_α and V_α are couplings satisfying (7.2). Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \lambda^{-1} (1 - e^{-\lambda}) \sum_\alpha p_\alpha E|U_\alpha - V_\alpha|.$$

As a quick check, if $\{X_\alpha, \alpha \in \Gamma\}$ is an independent family of indicators then we may select $U_\alpha = W$ and $V_\alpha = W_{(\alpha)}$ for each α . Then $U_\alpha - V_\alpha = X_\alpha \geq 0$ and $E|U_\alpha - V_\alpha| = E X_\alpha = p_\alpha$ leading to the theorem of Section 4 for the case of independent trials.

In general, we obtain a good Poisson approximation if the couplings (U_α, V_α) can be selected so that $E|U_\alpha - V_\alpha|$ is small. Roughly speaking, knowing $X_\alpha = 1$ should not affect the value of W by very much.

In many situations it is possible to elect $U_\alpha = W = X_\alpha + W_{(\alpha)}$ but the proper choice of coupling for V_α approaches an art form. The following generic formulation provides some guidance.

For each $\alpha \in \Gamma$, suppose $\{X'_{\alpha\beta}, \beta \in \Gamma\}$ is a family of ancillary random variables, defined on the same space, whose joint distribution is that of $(X_\beta, \beta \in \Gamma)$ conditioned on the event $X_\alpha = 1$. In notation: $\mathcal{L}(X'_{\alpha\beta}, \beta \in \Gamma) = \mathcal{L}(X_\beta, \beta \in \Gamma | X_\alpha = 1)$. Set $V_\alpha = \sum_{\beta \neq \alpha} X'_{\alpha\beta}$ and $U_\alpha = W = X_\alpha + \sum_{\beta \neq \alpha} X_\beta$. We have a bona fide coupling at hand with $\mathcal{L}(V_\alpha) = \mathcal{L}(W_{(\alpha)} | X_\alpha = 1)$ and $\mathcal{L}(U_\alpha) = \mathcal{L}(W)$.

Monotonicity relations between the coupling variables and the original variables helps simplify expressions further. For each α , let $\{\Gamma_\alpha^-, \Gamma_\alpha^0, \Gamma_\alpha^+\}$ be any partition of $\Gamma \setminus \{\alpha\}$ such that

$$X'_{\alpha\beta} \begin{cases} \leq X_\beta & \text{if } \beta \in \Gamma_\alpha^-, \\ \geq X_\beta & \text{if } \beta \in \Gamma_\alpha^+, \end{cases} \quad (7.3)$$

with there being no *a priori* monotone relationship between $X'_{\alpha\beta}$ and X_β for $\beta \in \Gamma_\alpha^0$. It is clear that such a partition is always possible for a given coupling (by the simple expedient of setting $\Gamma_\alpha^0 = \Gamma \setminus \{\alpha\}$) but to be effective we should attempt to select the ancillary variables $\{X'_{\alpha\beta}\}$ so that the set Γ_α^0 of "uncontrolled indices" is as small as possible. As each $X'_{\alpha\beta}$ is positive with probability one (why?) we may write

$$\begin{aligned} |U_\alpha - V_\alpha| &= \left| X_\alpha - \sum_{\beta \neq \alpha} (X'_{\alpha\beta} - X_\beta) \right| \\ &\leq X_\alpha - \sum_{\beta \in \Gamma_\alpha^-} (X'_{\alpha\beta} - X_\beta) + \sum_{\beta \in \Gamma_\alpha^0} (X'_{\alpha\beta} + X_\beta) + \sum_{\beta \in \Gamma_\alpha^+} (X'_{\alpha\beta} - X_\beta). \end{aligned} \quad (7.4)$$

Now, for each $\beta \neq \alpha$, it is clear by selection that $X'_{\alpha\beta}$ has the conditional distribution of X_β given that $X_\alpha = 1$. Accordingly,

$$p_\alpha E(X'_{\alpha\beta}) = P\{X_\alpha = 1\} E(X_\beta | X_\alpha = 1) = E(X_\alpha X_\beta) = \text{Cov}(X_\alpha, X_\beta) + p_\alpha p_\beta.$$

Now, in view of the monotone nature of the partition (7.3), we have

$$\text{Cov}(X_\alpha, X_\beta) = p_\alpha E(X'_{\alpha\beta} - X_\beta) \begin{cases} \leq 0 & \text{if } \beta \in \Gamma_\alpha^-, \\ \geq 0 & \text{if } \beta \in \Gamma_\alpha^+. \end{cases} \quad (7.5)$$

By taking expectations of both sides of (7.4) we may hence bound the coupling summands on the right in terms of covariances via

$$\begin{aligned} p_\alpha E|U_\alpha - V_\alpha| &\leq p_\alpha^2 - \sum_{\beta \in \Gamma_\alpha^-} \text{Cov}(X_\alpha, X_\beta) \\ &\quad + \sum_{\beta \in \Gamma_\alpha^0} [\text{Cov}(X_\alpha, X_\beta) + 2p_\alpha p_\beta] + \sum_{\beta \in \Gamma_\alpha^+} \text{Cov}(X_\alpha, X_\beta) \\ &= p_\alpha^2 + \sum_{\beta \notin \Gamma_\alpha^0} |\text{Cov}(X_\alpha, X_\beta)| + \sum_{\beta \in \Gamma_\alpha^0} \text{Cov}(X_\alpha, X_\beta) + 2 \sum_{\beta \in \Gamma_\alpha^0} p_\alpha p_\beta. \end{aligned}$$

The deviation from the Poisson may now be estimated from Theorem 1.

THEOREM 2 For each $\alpha \in \Gamma$ suppose that $\mathcal{L}(X'_{\alpha\beta}, \beta \in \Gamma) = \mathcal{L}(X_\beta, \beta \in \Gamma | X_\alpha = 1)$, and suppose $\{\Gamma_\alpha^-, \Gamma_\alpha^0, \Gamma_\alpha^+\}$ is a partition of $\Gamma \setminus \{\alpha\}$ satisfying (7.3). Then

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) &\leq \frac{1 - e^{-\lambda}}{\lambda} \left[\sum_{\alpha} p_{\alpha}^2 + \sum_{\alpha} \sum_{\beta \notin \Gamma_{\alpha}^0} |\text{Cov}(X_{\alpha}, X_{\beta})| \right. \\ &\quad \left. + \sum_{\alpha} \sum_{\beta \in \Gamma_{\alpha}^0} \text{Cov}(X_{\alpha}, X_{\beta}) + 2 \sum_{\alpha} \sum_{\beta \in \Gamma_{\alpha}^0} p_{\alpha} p_{\beta} \right]. \quad (7.6) \end{aligned}$$

The bound becomes particularly efficacious if we can select the coupling variables so that there is a one-way monotone relationship with only one of the bookend inequalities in (7.3) in play.

DEFINITION We say that the family of indicator variables $\{X_{\alpha}, \alpha \in \Gamma\}$ is *positively related* if there exists a family of coupling variables $\{X'_{\alpha\beta}\}$ such that, for each α , $X'_{\alpha\beta} \geq X_{\beta}$ for all $\beta \neq \alpha$. We say that $\{X_{\alpha}, \alpha \in \Gamma\}$ is *negatively related* if, for each α , $X'_{\alpha\beta} \leq X_{\beta}$ for all $\beta \neq \alpha$. We say that $\{X_{\alpha}\}$ is *monotonically related* if the variables are either positively or negatively related.

Thus, $\{X_{\alpha}\}$ is positively related if $\Gamma_{\alpha}^- = \Gamma_{\alpha}^0 = \emptyset$ and $\Gamma_{\alpha}^+ = \Gamma \setminus \{\alpha\}$ while $\{X_{\alpha}\}$ is negatively related if $\Gamma_{\alpha}^+ = \Gamma_{\alpha}^0 = \emptyset$ and $\Gamma_{\alpha}^- = \Gamma \setminus \{\alpha\}$. In either case, the latter two sums on the right in (7.6) vanish and the expression can be consolidated further in terms of the variance of W ; the situation becomes particularly transparent if the variables $\{X_{\alpha}, \alpha \in \Gamma\}$ are *exchangeable*, that is to say, their joint distribution is invariant with respect to permutations of coordinates.

COROLLARY Suppose the family of indicator variables $\{X_{\alpha}, \alpha \in \Gamma\}$ is monotonically related. Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \begin{cases} 1 - \frac{\text{Var}(W)}{\mathbb{E}(W)} & \text{if } \{X_{\alpha}\} \text{ is negatively related,} \\ \frac{1}{\mathbb{E}(W)} \left[2 \sum_{\alpha} p_{\alpha}^2 + \text{Var}(W) \right] - 1 & \text{if } \{X_{\alpha}\} \text{ is positively related.} \end{cases}$$

If, additionally, $\text{card } \Gamma = n$ and the variables $\{X_{\alpha}\}$ are exchangeable with $p_{\alpha} = p$ for all α and $\mathbb{E}(X_{\alpha} X_{\beta}) = \kappa$ for all $\beta \neq \alpha$, then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq p + (n - 1) \left| \frac{\kappa - p^2}{p} \right|.$$

PROOF: Suppose $\{X_{\alpha}\}$ is negatively related. Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \frac{1 - e^{-\lambda}}{\lambda} \left[\sum_{\alpha} p_{\alpha}^2 - \sum_{\alpha} \sum_{\beta \in \Gamma_{\alpha}^-} \text{Cov}(X_{\alpha}, X_{\beta}) \right]$$

which we may write in terms of the mean and the variance of W via

$$\begin{aligned}\text{Var}(W) &= \sum_{\alpha} \text{Var}(X_{\alpha}) + \sum_{\alpha} \sum_{\beta \neq \alpha} \text{Cov}(X_{\alpha}, X_{\beta}) \\ &= \sum_{\alpha} p_{\alpha} - \sum_{\alpha} p_{\alpha}^2 + \sum_{\alpha} \sum_{\beta \neq \alpha} \text{Cov}(X_{\alpha}, X_{\beta}).\end{aligned}$$

The argument follows similar lines for positively related variables. If, additionally, the variables $\{X_{\alpha}\}$ are exchangeable then $E(W) = np$ and $\text{Var}(W) = np - np^2 + n(n-1)(\kappa - p^2)$, and we may consolidate bounds via (7.5). ▶

Thus, Poisson approximation is good when the variance of W is approximately equal to its mean. (The reader will recall that the Poisson distribution has this basic property.) In particular, if $\{X_{\alpha}\}$ is monotonically related and exchangeable then Poisson approximation is good if $|\kappa - p^2| = o(p/n)$.

8 Matchings, ménages, permutations

Problems involving random permutations $\Pi: (1, \dots, n) \mapsto (\Pi_1, \dots, \Pi_n)$ provide fertile ground for the application of coupling ideas. Following Pólya's wise dictum, to settle ideas we start with the simplest, non-trivial version of the problem—an old favourite.

EXAMPLE 1) *Once more, le problème des rencontres.* For each $1 \leq j \leq n$, let A_j denote the collection of permutations (π_1, \dots, π_n) for which $\pi_j = j$. Let $X_j = 1_{A_j}(\Pi) = 1(\Pi_j = j)$ be the indicator for the event that a random permutation leaves index j in its original position, that is, there is a match at location j . In the setting of de Montmort's classical problem one is interested in the number of matches $W = \sum_{j=1}^n X_j$. The variables X_j in this setting are clearly exchangeable and we have $E(X_j) = p = 1/n$ for each j whence $E(W) = \sum_j E(X_j) = np = 1$. The calculation of the variance is almost as easy. If $j \neq k$ we have $E(X_j X_k) = \kappa = 1/n(n-1)$ so that $\text{Cov}(X_j, X_k) = \kappa - p^2 = 1/n^2(n-1) > 0$. It follows that

$$\begin{aligned}\text{Var}(W) &= \sum_j \text{Var}(X_j) + \sum_j \sum_{k \neq j} \text{Cov}(X_j, X_k) \\ &= n \cdot \frac{1}{n} \left(1 - \frac{1}{n}\right) + n(n-1) \cdot \frac{1}{n^2(n-1)} = 1 - \frac{1}{n} + \frac{1}{n} = 1.\end{aligned}$$

This is promising, the mean and the variance are equal. It now only needs the establishment of a coupling. Consider any index j and a random permutation $\Pi = (\Pi_1, \dots, \Pi_n)$. The analysis of Section II.4 now suggests that a simple transposition of elements to place index j at its original position may do the trick. Let κ_j denote the index for which $\Pi_{\kappa_j} = j$. Form the permutation $\Pi': (1, \dots, n) \mapsto (\Pi'_1, \dots, \Pi'_n)$ by transposing the pair of elements $\Pi_{\kappa_j} = j$ and Π_j in the original permutation: $\Pi'_j = j$, $\Pi'_{\kappa_j} = \Pi_j$, and $\Pi'_k = \Pi_k$ if $k \notin \{j, \kappa_j\}$. We now consider the

ancillary indicator variables $X'_k = 1_{A_k}(\Pi') = 1(\Pi'_k = k)$. As $X'_j = 1$ by the structure of the transposition, it is clear that $\mathcal{L}(X'_1, \dots, X'_n) = \mathcal{L}(X_1, \dots, X_n \mid X_j = 1)$ and we now have a coupling to work with. By the nature of the construction, $X'_k = X_k$ if $k \notin \{j, \kappa_j\}$, $X'_j = 1 \geq X_j$, and $X'_{\kappa_j} = 1(\Pi'_{\kappa_j} = \kappa_j) = 1(\Pi_j = \kappa_j) \geq 1(\Pi_{\kappa_j} = \kappa_j) = X_{\kappa_j}$ as $\Pi_{\kappa_j} = j$. It follows that $\{X_1, \dots, X_n\}$ is a family of positively related variables. The corollary of the previous section hence shows that

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq p + (n-1) \frac{\kappa - p^2}{p} = \frac{1}{n} + \frac{(n-1)n}{n^2(n-1)} = \frac{2}{n}.$$

While this is sufficient to show Poisson convergence as $n \rightarrow \infty$, the bound itself is not very sharp. Its value lies in the establishment of a framework within which more general permutation problems can be tackled.

A sharper bound. The elementary inclusion–exclusion argument leading to (IV.1.4) yields the point probabilities $P\{W = m\} = P_n(m) = \frac{1}{m!} \sum_{k=0}^{n-m} \frac{(-1)^k}{k!}$. The sum on the right is the truncation of the Taylor series for e^{-1} and so, writing $p(m; 1) = \frac{e^{-1}}{m!} = \frac{1}{m!} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}$ for the Poisson(1) point probabilities, we see that

$$|P_n(m) - p(m; 1)| = \frac{1}{m!} \left(\frac{1}{(n-m+1)!} - \frac{1}{(n-m+2)!} + \dots \right) \leq \frac{1}{m!(n-m+1)!}$$

as the terms of the series within round brackets are decreasing and alternating in sign. Robbins's lower bound for the factorial (XVI.6.1) shows that $k! > k^k e^{-k}$ for each k . And so, writing $f(x) = -x \log(x) - (n-x+1) \log(n-x+1)$ and bounding $m!$ and $(n-m+1)!$ from below separately, we obtain $(m!(n-m+1)!)^{-1} < \exp\{n+1+f(m)\}$. An easy differentiation shows that $f(x)$ is maximised for $x = (n+1)/2$ and so, by summing terms, we obtain

$$\begin{aligned} \sum_{m \leq n} |P_n(m) - p(m; 1)| &< \sum_{m \leq n} \exp\{n+1+f(m)\} \\ &\leq n \exp\left\{n+1+f\left(\frac{n+1}{2}\right)\right\} = n \left(\frac{2e}{n+1}\right)^{n+1} = \mathcal{O}\left\{\left(\frac{2e}{n}\right)^n\right\}. \end{aligned} \quad (8.1)$$

On the other hand, $P_n(m) = 0$ for $m > n$, whence

$$\sum_{m > n} |P_n(m) - p(m; 1)| = \sum_{m > n} \frac{e^{-1}}{m!} = \frac{e^{-1}}{(n+1)!} \left(1 + \frac{(n+1)!}{(n+2)!} + \frac{(n+1)!}{(n+3)!} + \dots\right).$$

By comparison of the expression within round brackets on the right with the terms of a geometric series with ratio $1/(n+1)$ we obtain

$$\sum_{m > n} |P_n(m) - p(m; 1)| \leq \frac{e^{-1}}{(n+1)! \left(1 - \frac{1}{n+1}\right)} = \frac{e^{-1}}{n \cdot n!} = \mathcal{O}\left\{\left(\frac{e}{n}\right)^{n+3/2}\right\}, \quad (8.2)$$

the bound on the right subdominant compared to (8.1). Combining the bounds for the left and right sums (8.1, 8.2), it follows that

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \sum_{m=0}^{\infty} |P_n(m) - p(m; 1)| = \mathcal{O}\left\{\left(\frac{2e}{n}\right)^n\right\}$$

and convergence to the Poisson is very rapid indeed, much more so than is suggested by the Stein–Chen method. ►

A slight recasting of the matching problem permits a profitable generalisation. Write A_{ij} for the set of permutations (π_1, \dots, π_n) for which $\pi_i = j$ and introduce the indicator variables $X_{ij} = 1_{A_{ij}}(\Pi) = 1(\Pi_i = j)$. Then $X = [X_{ij}]$ represents the permutation matrix of order n with unit entries at locations $(1, \Pi_1), \dots, (n, \Pi_n)$. If $S = [s_{ij}]$ represents the identity matrix of order n then the number of matches in the random permutation Π may be written in the form $W = \sum_i s_{i\Pi_i} = \sum_i \sum_j s_{ij} X_{ij}$. The point of the exercise is that the analysis of the matching problem may be essentially reproduced for *any matrix S whose elements are zeros and ones*.

Let $S = [s_{ij}]$ now be any square matrix of order n with entries $s_{ij} \in \{0, 1\}$. We think of S as a *selection matrix* which picks out favourable outcomes of the permutation. Let Γ represent the natural index set of the n^2 integer pairs (i, j) with $1 \leq i, j \leq n$ and let $\alpha = (i, j)$ represent a generic element of Γ . Write

$$W = \sum_{i=1}^n s_{i\Pi_i} = \sum_{i=1}^n \sum_{j=1}^n s_{ij} X_{ij} = \sum_{\alpha \in \Gamma} s_{\alpha} X_{\alpha}.$$

We are hence dealing with the family of indicators $\{s_{\alpha} X_{\alpha}, \alpha \in \Gamma\}$ with $W = \sum_{\alpha} s_{\alpha} X_{\alpha}$ representing the number of permutation indices selected by the matrix S .

A detour to pick up some new notation now simplifies exposition later. Write $s_{\bullet\bullet} = \sum_i \sum_j s_{ij}$ for the sum of terms of S , $s_{i\bullet} = \sum_j s_{ij}$ for the row sums, and $s_{\bullet j} = \sum_i s_{ij}$ for the column sums. For any $\alpha = (i, j)$, the indicator $s_{\alpha} X_{\alpha}$ has expectation $E(s_{\alpha} X_{\alpha}) = s_{\alpha} E(X_{\alpha}) = s_{\alpha}/n$ and so

$$\lambda = E(W) = \sum_{\alpha} E(s_{\alpha} X_{\alpha}) = \frac{s_{\bullet\bullet}}{n}. \quad (8.3)$$

For each $\alpha = (i, j) \in \Gamma$, introduce the sets $\Gamma_{\alpha}^- = \{(i, s) : s \neq j\} \cup \{(r, j) : r \neq i\}$ and $\Gamma_{\alpha}^+ = \{(r, s) : r \neq i, s \neq j\}$ which partition $\Gamma \setminus \{\alpha\}$. As each row and each column of the permutation matrix X contains a single one we have $E(X_{\alpha} X_{\beta}) = 0$ if $\beta \in \Gamma_{\alpha}^-$ while $E(X_{\alpha} X_{\beta}) = 1/n(n-1)$ if $\beta \in \Gamma_{\alpha}^+$. Thus,

$$\text{Cov}(X_{\alpha}, X_{\beta}) = E(X_{\alpha} X_{\beta}) - E(X_{\alpha}) E(X_{\beta}) = \begin{cases} -1/n^2 & \text{if } \beta \in \Gamma_{\alpha}^-, \\ 1/n^2(n-1) & \text{if } \beta \in \Gamma_{\alpha}^+. \end{cases}$$

Obtaining an expression for the variance in terms of the various row and column sums of the selection matrix S is now just a matter of algebra. We have

$$\begin{aligned} \text{Var}(W) &= \sum_{\alpha} s_{\alpha}^2 \text{Var}(X_{\alpha}) + \sum_{\alpha \neq \beta} s_{\alpha} s_{\beta} \text{Cov}(X_{\alpha}, X_{\beta}) \\ &= \frac{1}{n} \left(1 - \frac{1}{n}\right) \sum_{\alpha} s_{\alpha}^2 - \frac{1}{n^2} \sum_{\alpha} s_{\alpha} \sum_{\beta \in \Gamma_{\alpha}^-} s_{\beta} + \frac{1}{n^2(n-1)} \sum_{\alpha} s_{\alpha} \sum_{\beta \in \Gamma_{\alpha}^+} s_{\beta}. \end{aligned}$$

Representing Γ as a two-dimensional array of indices, we may identify Γ_α^- with the row and column containing the point $\alpha = (i, j)$ in the array with the point α itself excluded. We hence have $\sum_{\beta \in \Gamma_\alpha^-} s_\beta = s_{i\bullet} + s_{\bullet j} - 2s_{ij}$ and $\sum_{\beta \in \Gamma_\alpha^+} s_\beta = s_{\bullet\bullet} - s_{i\bullet} - s_{\bullet j} + s_{ij}$. (In the addition and subtraction of the term s_{ij} to massage the sums over Γ_α^- and Γ_α^+ into expressions involving $s_{\bullet\bullet}$, $s_{i\bullet}$, and $s_{\bullet j}$ the reader may recognise with pleasure an elementary inclusion and exclusion!) As $s_\alpha^2 = s_{ij}^2 = s_{ij}$, by summing over $\alpha = (i, j)$ we hence obtain

$$\begin{aligned} \text{Var}(W) &= \frac{1}{n} \left(1 - \frac{1}{n} \right) s_{\bullet\bullet} - \frac{1}{n^2} \sum_{i,j} s_{ij} (s_{i\bullet} + s_{\bullet j}) + \frac{2}{n^2} s_{\bullet\bullet} \\ &\quad - \frac{1}{n^2(n-1)} \sum_{i,j} s_{ij} (s_{i\bullet} + s_{\bullet j}) + \frac{1}{n^2(n-1)} (s_{\bullet\bullet}^2 + s_{\bullet\bullet}). \end{aligned}$$

By summing over i and j in turn we see that $\sum_{i,j} s_{ij} (s_{i\bullet} + s_{\bullet j}) = \sum_i s_{i\bullet}^2 + \sum_j s_{\bullet j}^2$. Pooling terms and simplifying we obtain the variance in terms of the row and column sums of S :

$$\text{Var}(W) = \frac{s_{\bullet\bullet}}{n} - \frac{1}{n(n-1)} \left(\sum_i s_{i\bullet}^2 + \sum_j s_{\bullet j}^2 - \frac{s_{\bullet\bullet}^2}{n} - s_{\bullet\bullet} \right). \quad (8.4)$$

The nature of the selection matrix S permits the discovery of a useful feature.

THEOREM 1 *If S is a permutation matrix or the zero matrix then $\text{Var}(W) = \mathbb{E}(W)$. In all other cases $\text{Var}(W) < \mathbb{E}(W)$.*

PROOF: As the entries of S are all either zero or one, we have

$$n \mathbb{E}(W) = s_{\bullet\bullet} = \sum_{i,j} s_{ij} = \sum_i s_{i\bullet} \leq \sum_i s_{i\bullet}^2.$$

Equality holds if, and only if, $s_{i\bullet} = 0$ or 1 for each i . Now, by the familiar Cauchy–Schwarz inequality for the usual inner product in \mathbb{R}^n , we have

$$n^2 \mathbb{E}(W)^2 = \left(\sum_{i,j} s_{ij} \right)^2 = \left(\sum_j 1 \cdot s_{\bullet j} \right)^2 \leq \left(\sum_j 1^2 \right) \left(\sum_j s_{\bullet j}^2 \right) = n \sum_j s_{\bullet j}^2,$$

equality again holding if, and only if, $s_{\bullet j} = 0$ or 1 for each j . The term in the round brackets on the right in (8.4) is hence positive and takes value zero if, and only if, S is either a permutation matrix or the zero matrix. ▶

In view of this result, we may anticipate Poisson behaviour if the selection matrix S is suitably sparse. All we need now is a reasonable coupling.

THE SWAP: A transposition of the permutation Π again leads to a useful coupling. Fix any element $\alpha = (i, j) \in \Gamma$ and let κ_j denote the index for which

$\Pi_{\kappa_j} = j$. Introduce the ancillary permutation $\Pi': (1, \dots, n) \mapsto (\Pi'_1, \dots, \Pi'_n)$ by swapping Π_i and Π_{κ_j} , and keeping the other indices of the permutation intact: formally, set $\Pi'_i = j$, $\Pi_{\kappa_j} = \Pi_i$, and $\Pi'_k = \Pi_k$ if $k \notin \{i, \kappa_j\}$. The permutation matrix $X = [X_\beta]_{\beta \in \Gamma}$ then transposes to a new permutation matrix $X' = [X'_{\alpha\beta}]_{\beta \in \Gamma}$ where, for each $\beta = (k, l)$, $X'_{\alpha\beta} = 1_{A_{kl}}(\Pi') = 1(\Pi'_k = l)$. The reader may well prefer the adjoining Figure 1 to the formalism.

If $X_\alpha = 1$ then the original permutation has $\Pi_i = j$ and no swap is needed. It is clear hence that $\mathcal{L}(X'_{\alpha\beta}, \beta \in \Gamma) = \mathcal{L}(X_\beta, \beta \in \Gamma \mid X_\alpha = 1)$ and we have a bona fide coupling in hand. Suppose now that $X_\alpha \neq 1$. Then $X'_{\alpha\beta} = X_\beta$ if $\beta \notin \{(i, \Pi_i), (\kappa_j, \Pi_i), (i, j), (\kappa_j, j)\}$ as the swap only involves these four excluded points. Of these points, (i, Π_i) and (κ_j, j) are in Γ_α^- , (κ_j, Π_i) is in Γ_α^+ , and $(i, j) = \alpha \notin \Gamma_\alpha^- \cup \Gamma_\alpha^+$. Introducing commas in the subscript to visually separate the indices corresponding to α and β , we have $X'_{ij, i\Pi_i} = 0 \leq X_{i\Pi_i}$ and $X'_{ij, \kappa_j j} = 0 \leq 1 = X_{\kappa_j j}$ and it follows that $X'_{\alpha\beta} \leq X_\beta$ if $\beta \in \Gamma_\alpha^-$. Finally, we verify that $X'_{ij, \kappa_j \Pi_i} = 1 \geq X_{\kappa_j \Pi_i}$ and so

$X'_{\alpha\beta} \geq X_\beta$ for $\beta \in \Gamma_\alpha^+$. We are now primed to use Theorem 7.2 with $\Gamma_\alpha^0 = \emptyset$ and the indicators X_α replaced by $s_\alpha X_\alpha$.

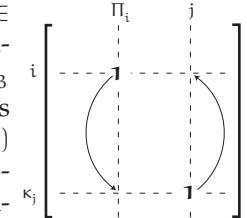


Figure 1: The swap.

THEOREM 2 Suppose $\Pi: (1, \dots, n) \mapsto (\Pi_1, \dots, \Pi_n)$ is a random permutation, $X = [X_{ij}]$ the corresponding permutation matrix with entries $X_{ij} = 1(\Pi_i = j)$, and $S = [s_{ij}]$ any selection matrix of order n . Let $W = \sum_i \sum_j s_{ij} X_{ij}$. Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq (1 - e^{-\mathbb{E}(W)}) \left[\frac{2\mathbb{E}(W)}{n} + \frac{n-2}{n} \left(1 - \frac{\text{Var}(W)}{\mathbb{E}(W)} \right) \right].$$

PROOF: With X_α replaced by $s_\alpha X_\alpha$, we begin with the expression in the square brackets on the right in (7.6). Only the first two terms contribute as $\Gamma_\alpha^0 = \emptyset$ and now the calculations mirror those for the variance of W , the only difference being a change in the sign of the sum over Γ_α^- . We have

$$\begin{aligned} \sum_\alpha \sum_{\beta \notin \Gamma_\alpha^0} |\text{Cov}(s_\alpha X_\alpha, s_\beta X_\beta)| &= \frac{1}{n^2} \sum_\alpha s_\alpha \sum_{\beta \in \Gamma_\alpha^-} s_\beta + \frac{1}{n^2(n-1)} \sum_\alpha s_\alpha \sum_{\beta \in \Gamma_\alpha^+} s_\beta \\ &= \frac{n-2}{n^2(n-1)} \left(\sum_i s_{i\bullet}^2 + \sum_j s_{\bullet j}^2 \right) - \frac{2s_{\bullet\bullet}}{n^2} + \frac{s_{\bullet\bullet}^2 + s_{\bullet\bullet}}{n^2(n-1)}. \end{aligned}$$

Identifying $p_\alpha^2 = \mathbb{E}(s_\alpha X_\alpha)^2 = s_\alpha^2/n^2 = s_\alpha/n^2$ in (7.6), we obtain

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq \frac{1 - e^{-s_{\bullet\bullet}/n}}{n(n-1)} \left[\frac{n-2}{s_{\bullet\bullet}} \left(\sum_i s_{i\bullet}^2 + \sum_j s_{\bullet j}^2 \right) + s_{\bullet\bullet} - n + 2 \right].$$

By (8.3), $s_{\bullet\bullet} = n \mathbb{E}(W)$, and, likewise, we may turn (8.4) to account and write $\sum_i s_{i\bullet}^2 + \sum_j s_{\bullet j}^2$ in terms of the mean and variance of W . Substitution completes the proof. ▶

This result was first proved by Chen.⁸ The usual culprits serve up illustrative uses of the technique.

EXAMPLES: 2) *Rencontres redux*. The selection matrix S is the identity matrix resulting in $E(W) = \text{Var}(W) = 1$. Consequently, $d_{\text{TV}}(\mathcal{L}(W), \mathcal{L}(Z)) \leq (1 - e^{-1})2/n$, and we recover the bound of Example 1.

3) *The game of Treize*. A standard pack of 52 cards is shuffled and 13 cards drawn in sequence. We identify Ace with 1 and Jack, Queen, and King with 11, 12, and 13, respectively, regardless of suit. We say that there is a match on draw number j if the face value of the j th card drawn is j . Let W be the number of matches.

The selection matrix S has order 52. Each of the first 13 rows of S have exactly four ones apiece, all subsequent rows comprised only of zeros. Each of the 52 columns of S have a single one. Thus, $s_{\bullet\bullet} = 13 \times 4 = 52$, $s_{i\bullet} = 4$ for $1 \leq i \leq 13$ and $s_{i\bullet} = 0$ for $14 \leq i \leq 52$, and $s_{\bullet j} = 1$ for each j . It follows that $E(W) = 1$ and $\text{Var}(W) = 16/17$ and substitution shows that

$$d_{\text{TV}}(\mathcal{L}(W), \mathcal{L}(Z)) \leq (1 - e^{-1}) \left[\frac{2}{52} + \frac{50}{52} \left(1 - \frac{16}{17}\right) \right] \approx 0.0601.$$

The first reported investigation of this game is by de Montmort in 1708. In 1711 Nicolas Bernoulli computed the exact value of the probability of obtaining no matches,

$$\mathbf{P}\{W = 0\} = \frac{1}{2} \left[1 - \frac{7672980411874433}{26816424180170625} \right] \approx 0.3569,$$

which differs from the Poisson point probability e^{-1} by 0.0109 to four decimal places.⁹ While the reader may well marvel at the prodigious care and precision that Bernoulli's hand computation must have entailed, the exact computation of these probabilities on a modern digital computer requires no great finesse. A quick calculation shows that the total variation distance is approximately 0.0109, the Stein–Chen bound exceeding the true value by about 5%.

4) *Le problème des ménages*. Posed in its classical form, there are n couples seated at a circular table, men and women alternating. A match is uncovered whenever a wife is seated next to her husband, W being the total number of matches. The problem posed by E. Lucas asks for the number of arrangements so that no husband is seated next to his wife; in our notation this number is given by $n! \mathbf{P}\{W = 0\}$.

⁸L. H. Y. Chen, "An approximation theorem for sums of certain randomly selected indicators", *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 33, pp. 69–74, 1975.

⁹P. R. de Montmort, *Essay d'analyse sur les jeux de hazard*. Paris: chez Jacque Quillau, 1708. N. Bernoulli's computation is reported in the second edition of de Montmort's work, *Revue & augmentée de plusieurs Lettres*. Paris: chez Jacque Quillau, 1713.

By symmetry, we may as well fix the seating of, say, the gentlemen in some order, say $1, \dots, n$ going clockwise, and then distribute the ladies randomly in the interstitial seating. The selection matrix S has order n and has a banded form with two contiguous ones in each row, and two contiguous ones in each column (1 and n are naturally considered to be contiguous in a circular arrangement). We thus have $s_{\bullet\bullet} = 2n$ and $s_{i\bullet} = s_{\bullet j} = 2$ for each i and j . It follows that $E(W) = 2$ and $\text{Var}(W) = 2(n-2)/(n-1)$, whence

$$d_{\text{TV}}(\mathcal{L}(W), \mathcal{L}(Z)) \leq (1 - e^{-2}) \left[\frac{4}{n} + \frac{n-2}{n} \left(1 - \frac{n-2}{n-1} \right) \right] = (1 - e^{-2}) \left[\frac{5}{n} - \frac{1}{n(n-1)} \right].$$

The exact form of the law of W is known in this case; see Problem IV.20. ►

Selection arrays with more than two indices can be handled by similar methods. The following example provides an illustrative setting.

EXAMPLE 5) Knox's leukaemia data.¹⁰ Knox's data indexes 96 cases of childhood leukaemia as well as spatial and temporal correlations: a pair (i, j) of children is flagged if the children lived near each other or if the presentations of their illnesses occurred close in time to each other. If the actual number of pairs for which the presentations of the disease were in proximity both in space and time is significantly larger than the number that could be expected if, say, the spatial coordinates were randomly permuted then one could conclude that there is a statistically significant association between the place and time of presentation. The connection with Poisson approximation lies in the idea that if the spatial or temporal coordinate is randomly permuted then, per our slogan, the number of coincidences in space and time should be governed roughly by a Poisson law. An analysis of Knox's data of this type was undertaken by Barbour and Eagleson to conclude that there is significant evidence for such a spatio-temporal association of disease presentation.¹¹ ►

The framework in which one can view higher-order arrays such as those governing Knox's data is as follows. Suppose $S = [s_{ij}]$ and $T = [t_{ij}]$ are two *symmetric* selection matrices of order n . We may view S and T as incidence matrices labelling the edges of two undirected graphs on n vertices. In Knox's data S could represent spatial correlations with $s_{ij} = 1$ when patients i and j lived near each other, the matrix T , likewise, could represent temporal coincidences in disease presentation. Let $\Pi: (1, \dots, n) \mapsto (\Pi_1, \dots, \Pi_n)$ represent a random permutation of, say, the spatial coordinate. Then $W = \sum_{i < j} s_{\Pi_i \Pi_j} t_{ij}$ counts the number of edges common to both graphs when the vertices of one are randomly permuted. We anticipate naturally enough that W should be accurately described by a Poisson law if the selection matrices S and T are suitably

¹⁰G. Knox, "Epidemiology of childhood leukaemia in Northumberland and Durham", *British Journal of Preventive Social Medicine*, vol. 18, pp. 17–24, 1964.

¹¹A. D. Barbour and G. K. Eagleson, "Poisson approximation for some statistics based on exchangeable trials", *Advances in Applied Probability*, vol. 15, pp. 585–600, 1983.

sparse. The methods of this chapter can be turned to account in this setting as well though the algebra does get inevitably messier; the interested reader will find the details fleshed out in the monograph of Barbour, Holst, and Janson.¹² When turned to the analysis of Knox's data they present excellent bounds on Poisson approximation but report regretfully that a definitive reconstruction is impossible as Knox's original data were destroyed in the course of a burglary.

9 Spacings and mosaics

Suppose ξ_1, \dots, ξ_n are random points in the unit interval $[0, 1]$. (We write ξ instead of X only because we would like to reserve the latter for indicators as customary in this chapter.) We suppose as usual that these points are generated by independent sampling from the uniform distribution. Let L_1, \dots, L_{n+1} be the successive spacings between these points. As the reader has seen in Section IX.2, it is cleanest to think of the random points as points on the circle \mathbb{T} of unit circumference obtained by rolling up the unit interval, identifying the points 0 and 1, and introducing an extra point $\xi_0 = 0$ to play the rôle of the origin. Starting with the point ξ_0 and enumerating clockwise around the circle for definiteness, the spacings L_1, \dots, L_{n+1} then partition \mathbb{T} into arcs of these lengths. In this view, the j th spacing L_j is the separation between the $(j-1)$ th and j th points encountered by starting at ξ_0 and proceeding clockwise.

Suppose $0 < r < 1$. We say that there is a *gap of size r* at the j th spacing if $L_j > r$. Let $X_j = 1_{(r,1)}(L_j) = 1(L_j > r)$ be the corresponding gap indicator. Then $W = \sum_{j=1}^{n+1} X_j$ represents the number of gaps. We may anticipate per our slogan that W will have an approximately Poisson distribution if r and n are suitably related.

It is natural to begin by considering the distribution of the spacings. These are given by de Finetti's theorem of Section IX.2. In particular, specialising (IX.2.2) to the case of the gap indicators $X_j = 1(L_j > r)$, we have

$$\mathbb{E}(X_j) = \mathbb{P}\{L_j > r\} = (1-r)_+^n \text{ and } \text{Var}(X_j) = (1-r)_+^n (1 - (1-r)_+^n). \quad (9.1)$$

Likewise, $X_j X_k = 1(L_j > r, L_k > r)$ is the indicator for the joint occurrence of gaps at positions j and k and (IX.2.3) then shows that

$$\mathbb{E}(X_j X_k) = (1-2r)_+^n, \text{ whence } \text{Cov}(X_j, X_k) = (1-2r)_+^n - (1-r)_+^{2n}. \quad (9.2)$$

As $(1-r)_+^{2n} = (1-2r+r^2)_+^n \geq (1-2r)_+^n$, the covariances are all negative and this suggests that the gap variables X_j may be negatively related. De Finetti's theorem now suggests how an appropriate coupling may be discovered.

¹²A. D. Barbour, L. Holst, and S. Janson, *op. cit.*

THE TINY SHIFT: If $0 < r < 1$ then $(1 - r)_+ = 1 - r$ is strictly positive and by conditioning, say, on the first gap, we obtain

$$\begin{aligned} \mathbf{P}\{L_2 > x_2, \dots, L_{n+1} > x_{n+1} \mid L_1 > r\} &= \frac{\mathbf{P}\{L_1 > r, L_2 > x_2, \dots, L_{n+1} > x_{n+1}\}}{\mathbf{P}\{L_1 > r\}} \\ &= \frac{(1 - r - x_2 - \dots - x_{n+1})_+^n}{(1 - r)^n} = \left(1 - \frac{x_2}{1 - r} - \dots - \frac{x_n}{1 - r}\right)_+^n. \end{aligned} \quad (9.3)$$

By comparison with (IX.2.1), if we identify x_1 with 0 and x_j with $x_j/(1 - r)$ for $2 \leq j \leq n + 1$, we recognise in the expression on the right just the unconditional probability $\mathbf{P}\{(1 - r)L_2 > x_2, \dots, (1 - r)L_{n+1} > x_{n+1}\}$. Of course, by exchangeability, we can repeat the argument by conditioning on any L_j instead of on L_1 and so, by symmetry, it suffices to exhibit a coupling for $j = 1$. A coupling now falls into our laps courtesy the relation (9.3). Set $X'_{11} = X_1 = 1(L_1 > r)$ and $X'_{1k} = 1((1 - r)L_k > r)$ for $k \neq 1$. At its heart the coupling shifts all points excepting one by a small amount. Then (9.3) shows that $\mathcal{L}(X'_{1k}, 2 \leq k \leq n + 1) = \mathcal{L}(X_k, 2 \leq k \leq n + 1 \mid X_1 = 1)$ and as it is trite that $X'_{1k} \leq X_k$ for each k , the gap indicators X_1, \dots, X_n are indeed negatively related.

In view of (9.1,9.2) we have

$$\begin{aligned} \mathbf{E}(W) &= (n + 1)(1 - r)_+^n, \\ \text{Var}(W) &= (n + 1)(1 - r)_+^n (1 - (1 - r)_+^n) + n(n + 1)((1 - 2r)_+^n - (1 - r)_+^{2n}). \end{aligned}$$

As the summands are negatively related we may now freely invoke the corollary of Section 7.

THEOREM Suppose W is the number of gaps of size r between $n + 1$ random points on the circle \mathbb{T} and suppose $0 < r < 1/2$. Let Z be a Poisson variable with mean $(n + 1)(1 - r)_+^n$. Then

$$d_{TV}(\mathcal{L}(W), \mathcal{L}(Z)) \leq n(1 - r)^n \left[1 - \frac{(1 - 2r)_+^n}{(1 - r)^{2n}} \right] + (1 - r)^n.$$

COROLLARY Suppose $r = r_n = n^{-1}(\log n + c + \eta_n)$ where $\eta_n \rightarrow 0$ as $n \rightarrow \infty$. Then the number W of gaps is governed asymptotically by the Poisson distribution with mean e^{-c} .

PROOF: Allowing $n \rightarrow \infty$, simple algebra shows that $(1 - r_n)^n < e^{-nr_n} = \mathcal{O}(n^{-1})$ and an only slightly more elaborate analysis shows

$$\begin{aligned} n(1 - r_n)^n &= \exp(\log n + n \log(1 - r_n)) \\ &= \exp(\log n - nr_n + \mathcal{O}(nr_n^2)) = \exp(-c - \eta_n + \mathcal{O}(n^{-1} \log n)) \rightarrow e^{-c}. \end{aligned}$$

Moreover, as $r_n < 1/2$, eventually, we have $r_n^2/(1 - r_n)^2 < 4r_n^2$ so that

$$\begin{aligned} 1 - \frac{(1 - 2r_n)^n}{(1 - r_n)^{2n}} &= 1 - \left(\frac{1 - 2r_n}{1 - 2r_n + r_n^2} \right)^n = 1 - \left(1 - \frac{r_n^2}{(1 - r_n)^2} \right)^n \\ &\leq 1 - (1 - 4r_n^2)^n = 4nr_n^2 + \mathcal{O}(nr_n^4) = \mathcal{O}(n^{-1} \log(n)^2). \end{aligned}$$

It follows that $d_{\text{TV}}(\mathcal{L}(W), \mathcal{L}(Z)) = \mathcal{O}(n^{-1} \log(n)^2)$ and the bound tends to zero as $n \rightarrow \infty$. ▶

This problem has received sustained attention over the years, in part no doubt because of its clean and simple formulation. The usage of the Stein–Chen method in this context dates to A. D. Barbour and L. Holst who provided explicit convergence rates.¹³

EXAMPLES: 1) *Coverage in the arc mosaic process.* Suppose one throws down a family of arcs of length r centred at the random points $\xi_0, \xi_1, \dots, \xi_n$ on the circle \mathbb{T} . Write \mathbb{A}_j for the arc centred at the point ξ_j . The collection of random arcs $\{\mathbb{A}_0, \mathbb{A}_1, \dots, \mathbb{A}_n\}$ covers a portion of the circle \mathbb{T} and forms the *simplest mosaic process*.

If $W = 0$ then there are no gaps and the mosaic process covers the circle. Under the conditions of the corollary, it follows that $P\{W = 0\} \rightarrow e^{-e^{-c}}$. This probability is close to one if the constant c is even slightly less than zero while it is close to zero if c is even slightly bigger than zero. Thus, a threshold function emerges for the property that the arc mosaic process covers the circle as the arc length $r = r_n$ moves past the order of $n^{-1} \log n$.

2) *Connectivity in a sensor network.* Given $0 < r < 1/2$, the selection of random points on the circle induces a *random geometric graph* \mathcal{G} with vertices at $\xi_0, \xi_1, \dots, \xi_n$ and an edge between two vertices ξ_j and ξ_k if, and only if, $|\xi_j - \xi_k| \leq r$. This graph will be connected if, and only if, there are either no gaps or there is precisely one gap; alternatively, the graph \mathcal{G} will consist of two or more separated components if there are two or more gaps. As $P\{W \leq 1\} \rightarrow e^{-e^{-c}}(1 + e^{-c})$, it follows that a threshold function for graph connectivity is also manifest when the arc length $r = r_n$ is of the order of $n^{-1} \log n$.

This type of result has use in *ad hoc* wireless networks of sensors. Suppose agents deployed randomly on the circle \mathbb{T} can communicate with other agents in their immediate proximity at a distance up to r along the circle. In this context one can think of r as representing the available transmission power. The induced random geometric graph then represents the sensor communication network and it is clearly a desirable feature that information from any agent can be relayed to any other agent, that is, the network is connected. Placed in

¹³A. D. Barbour and L. Holst, “Some applications of the Stein–Chen method for proving Poisson convergence”, *Advances in Applied Probability*, vol. 21, pp. 74–90, 1989.

this context our result says that for the network to be connected the available communication distance should scale at least as the order of $n^{-1} \log n$ where $n + 1$ is the number of randomly deployed sensors. ▶

The algebra obscures to some extent the simple idea behind the construction of the coupling. What is at its heart? It is the observation that the variables $M_2 = \frac{L_2}{1-L_1}, \dots, M_{n+1} = \frac{L_{n+1}}{1-L_1}$ represent the spacings of $n - 1$ random points in the unit interval and are *independent* of L_1 . Introduce notation $F(x) = P\{L_1 \leq x\}$ and $G(x) = P\{L_1 \leq x \mid L_1 > r\}$ for the unconditional and conditional marginal d.f.s of L_1 , respectively. Then $U_1 = F(L_1)$ is uniformly distributed in the unit interval and $L'_1 = G^{-1}(F(L_1))$ is hence a variable with distribution G that is determined purely by L_1 . Form the variables $L'_2 = (1 - L'_1)M_2, \dots, L'_{n+1} = (1 - L'_1)M_{n+1}$. The following intuitive and basic result which the reader is cordially invited to verify follows now because M_2, \dots, M_{n+1} are independent of L_1 .

$$\text{LEMMA } \mathcal{L}(L'_2, \dots, L'_{n+1}) = \mathcal{L}(L_2, \dots, L_{n+1} \mid L_1 > r).$$

This understanding allows us to construct couplings in other problems of this stripe; see the *Problems*.

The setting carries over to two and more dimensions though we now have to determine what exactly is meant by a spacing. Suppose ξ_1, \dots, ξ_n is obtained by independent selection from the uniform distribution in the unit cube $[-\frac{1}{2}, \frac{1}{2}]^d$ in d dimensions.¹⁴ When $d > 1$ the circular order of the points which determines the spacings is lost but we can still parlay the basic property that the spacings identify neighbouring points. A natural modification of the notion of proximity in dimensions greater than one is then to consider the nearest-neighbour distances $S_j = \min_{k \neq j} \|\xi_k - \xi_j\|$ where $\|\cdot\|$ represents the Euclidean norm (or even a generic L^p norm) in \mathbb{R}^d . Gaps are now replaced in two and more dimensions by an idea of geometric isolation. We write $X_j = 1(S_j > r)$ for the indicator of the event that point ξ_j is separated by a distance of at least r from all the other points and say that ξ_j is *isolated* if $X_j = 1$. The setting may be visualised by throwing down balls of radius r at random, the j th ball centred at the random point ξ_j . This forms a mosaic process in the unit cube.¹⁵ Figure 2 shows a mosaic in the unit square in two dimensions.

¹⁴The positioning of the cube makes little matter but it is algebraically more convenient to centre it at the origin with faces parallel to the axes.

¹⁵There is an extensive literature on the subject initiated by Gilbert's paper in 1961: E. N. Gilbert, "Random plane networks", *Journal for the Society of Industrial and Applied Mathematics*, vol. 9, pp. 533–553, 1961. In the general theory it is more satisfactory to throw down the balls or, more generally, random "smallish" sets at Poisson points in Euclidean space. Restricted to, say, the unit cube the model is approximately ours. A detailed study may be found in P. Hall, *Introduction to the Theory of Coverage Processes*. New York: John Wiley, 1988.

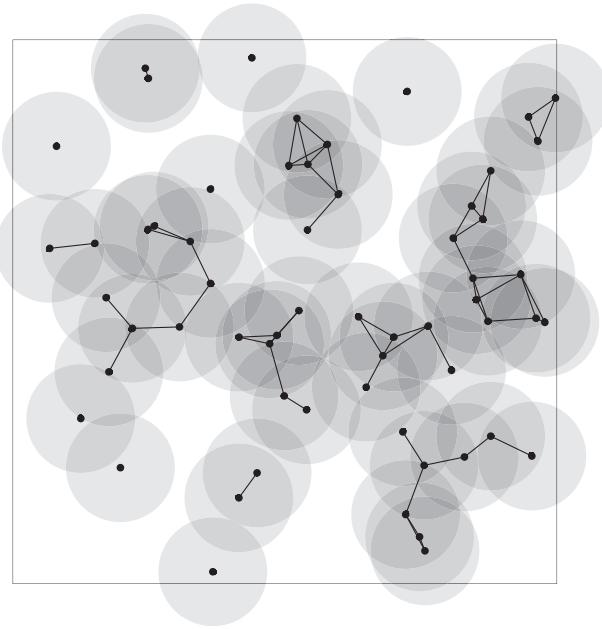
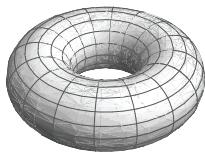


Figure 2: A mosaic of 60 randomly placed discs of radius $r = 0.1$ in the unit square. The mosaic process induces a geometric graph on vertices at the disc centres with edges connecting vertices which are in proximity ($\text{distance} \leq 0.1$).

From an analytical point of view the choice of cube (ball or other convex regular region) is not ideal because the faces of the cube (and, more generally, the boundaries of the chosen region) create awkward edge effects. These effects are negligible in two dimensions (if r is small)—though the effort needed to show it is significant. Boundary effects become increasingly unpleasant and onerous in dimensions three onwards and can no longer be neglected.

It is preferable hence to avoid the issue of the boundary altogether by working on the d -dimensional torus \mathbb{T}^d obtained by identifying the opposite faces of the cube $[-\frac{1}{2}, \frac{1}{2}]^d$ and equipping the resulting surface with the natural toroidal metric $\rho(\mathbf{x}, \mathbf{y}) = \min_{z \in \mathbb{Z}^d} \|\mathbf{x} - \mathbf{y} + z\|$ which identifies points if they are separated by integer coordinate shifts. The nearest-neighbour distances with respect to the toroidal metric now become $S_j = \min_{k \neq j} \rho(\xi_k, \xi_j)$. A visualisation of the two-dimensional torus by extrusion into the third dimension is shown in Figure 3.¹⁶

¹⁶One may choose instead to work on the surface \mathbb{S}^d of the unit ball in $d+1$ dimensions, distances mapped by the natural geodesic metric, the choice of manifold again avoiding boundary issues. The character of the results remains the same.



Write $W = X_1 + \dots + X_n$ for the number of isolated points on the d -dimensional torus. By Theorem XIV.6.1 we may identify $V_d = \pi^{d/2} / \Gamma(\frac{d}{2} + 1)$ with the volume of the unit ball in d dimensions and so the probability that a randomly selected point on the torus falls in the ball of radius r centred at the point ξ_1 is given by $V_d r^d$. It follows that

$$E(X_1) = P\{S_1 > r\} = (1 - V_d r^d)^{n-1}.$$

As the variables X_1, \dots, X_n are clearly exchangeable, we have $E(W) = n E(X_1)$. We anticipate that if r is suitably small then W will be approximately Poisson. In particular, if $r = r_n = \sqrt[d]{(\log n + c + \eta_n)/(nV_d)}$ where $\eta_n = \mathcal{O}\left(\frac{\log n}{n}\right)$, then $E(W) \rightarrow e^{-c}$ and we anticipate, in analogy with the corollary of this section for the one-dimensional setting, that W behaves asymptotically like a Poisson variable with mean e^{-c} . As the dependencies across the variables are pervasive, but weak, it is natural to attempt to verify this by constructing a coupling. If the reader attempts this along the lines followed in one dimension she will find to her chagrin that the circular order in one dimension played a critical rôle in the construction of the coupling: the basic idea behind the coupling is that the original set of points is replaced by a new set of points obtained by shifting all but one of the points a small distance along the circle. With no natural notion of order in dimensions two onwards she may feel compelled to abandon this approach. How now to proceed?

THE BIG SHIFT: As conditioning on the event $S_1 > r$ eliminates points in close proximity to ξ_1 , a natural idea is to move those points in the vicinity of ξ_1 randomly elsewhere on the torus. To make this idea concrete, let v_2, \dots, v_n be a random set of points, independent of ξ_2, \dots, ξ_n , chosen by independent selection from the uniform distribution on the set of points $\mathbb{T}^d \setminus \{v : \rho(v, \xi_1) \leq r\}$ consisting of the torus with the ball of radius r centred at ξ_1 expurgated. We now form the auxiliary set of random points ξ'_2, \dots, ξ'_n where $\xi'_k = \xi_k$ if $\rho(\xi_k, \xi_1) > r$ and $\xi'_k = v_k$ if $\rho(\xi_k, \xi_1) \leq r$, and, setting $S'_{1k} = \min_{i \neq k} \rho(\xi'_i, \xi'_k)$, form the proximity indicator variables $X'_{1k} = 1(S'_{1k} > r)$. It is clear from the construction that $V_1 = X'_{12} + \dots + X'_{1n}$ has the distribution of $W_{(1)} = X_2 + \dots + X_n$ conditioned on $S_1 > r$ and so, with $U_1 = W = X_1 + \dots + X_n$, we have a bona fide coupling (U_1, V_1) . Poisson approximation now carries through though the level of algebraic detail rapidly becomes tiresome; I will simply leave the reader with pointers to the literature if she wishes to explore further.¹⁷

¹⁷The coupling shown here is fleshed out in the previously cited monograph of Barbour, Holst, and Janson but other approaches are possible. Percolation methods are used in P. Gupta and P. R. Kumar, "Critical power for asymptotic connectivity in wireless networks", in *Stochastic Analysis, Control, Optimization and Applications: A Volume in Honor of W. H. Fleming* (eds W. M. McNeany, G. Yin, and Q. Zhang). Boston: Birkhäuser, pp. 547–566, 1998. A slew of results on the topic exploiting

10 Problems

1. *Paths in $G_{n,p}$.* We say that there is a path of length k in a graph if there exist vertices v_0, v_1, \dots, v_k with edges between adjacent vertices v_j and v_{j+1} . Provide an estimate for the accuracy of Poisson approximation for the number of paths of length k .

2. *Continuation, threshold function.* Show that if $p = p_n \ll n^{-(k+1)/k}$ then there are asymptotically no paths of length k in $G_{n,p}$ and that if $p = p_n \gg n^{-(k+1)/k}$ then there exist (many) paths of length k in $G_{n,p}$. In other words, show that $n^{-(k+1)/k}$ is a threshold function for the attribute that a path of length k exists in $G_{n,p}$.

3. *Paths in a random geometric graph.* Select n random points in the torus formed from the unit square $[-\frac{1}{2}, \frac{1}{2}]^2$ by identifying the left and right edges as well as the bottom and top edges. Suppose $0 < r < 1/2$ and form the random geometric graph with vertices at the random points and edges between vertices separated by a toroidal distance of no more than r . Determine the accuracy of Poisson approximation for the number of paths of length k in the induced random geometric graph. Thence, determine a threshold function $r = r_n$ for this property.

4. *Multiple matchings.* Two identical decks of n distinct cards are each matched against a target deck. Determine the probability $Q_n(m)$ that there are exactly m double matches. Show that $Q_n(0) \rightarrow 1$ as $n \rightarrow \infty$.

5. *Mixed multiple matchings.* Modify the matching procedure of the previous problem as follows. Shuffle the two packs together and out of the $2n$ -card composite pack select n cards at random. Match these n cards against the target deck. Determine the probability $Q'_n(0)$ that there are no matches and show that $Q'_n(0) \rightarrow e^{-1}$ as $n \rightarrow \infty$. Matchings of this type were considered by de Montmort.

6. *The game of Treize.* A pack of $N = rn$ cards comprised of r suits, each suit comprised of cards of face values 1 through n is shuffled and n cards drawn. A match occurs at draw number j if the j th card drawn has face value j , irrespective of suit. Let W be the number of matches. Determine $\lambda = E(W)$ and $\text{Var}(W)$. Thence bound the total variation distance between the law of W and the Poisson distribution of mean λ .

7. *Continuation.* Repeat the calculations in the setting of the previous problem when all N cards are drawn. A card of face value j gives a match if it occurs in the draws $j, j+n, \dots, j+(r-1)n$ and W is again the total number of matches.

8. *The general matching problem.* Let A and B be two decks of n cards apiece. Each deck has cards of r face values with A and B having a_j and b_j cards, respectively, of face value j , $1 \leq j \leq r$. The two decks are paired at random, each card from A paired with a card from B , all $n!$ pairings having equal probability. A pair forms a match if the card from A and the paired card from B have the same face value. Determine a total variation bound on the goodness of fit of an appropriately chosen Poisson distribution to the law of W , the total number of matches.

the local approach in the Stein–Chen method may be found in the monograph by M. Penrose, *Random Geometric Graphs*. Oxford: Oxford University Press, 2003. Kunniyur and Venkatesh use classical sieve methods to show Poisson approximation in the setting of the unit disc rather than the unit square: S. Kunniyur and S. S. Venkatesh, “Threshold functions, node isolation, and emergent lacunae in sensor networks”, *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5352–5372, 2006. The difficulties with the boundary are in clear evidence in this paper.

9. *Approximate matches and ménage problems.* It is convenient in this problem to begin counts at 0 instead of 1 and to identify n with 0. Let $\Pi: (0, 1, \dots, n-1) \mapsto (\Pi_0, \Pi_1, \dots, \Pi_{n-1})$ be a random permutation. With addition in subscripts interpreted modulo n , say that an *approximate match* occurs at position j if any of $\Pi_j, \Pi_{j+1}, \dots, \Pi_{j+r-1}$ is equal to j . Here $1 \leq r \leq n$ is a fixed positive integer. Evaluate the quality of a Poisson approximation for W in terms of r and n . The case $r = 1$ corresponds to de Montmort's problème des rencontres; $r = 2$ corresponds to the classical problème des ménages of Lucas.

10. *From the exponential distribution to spacings.* Let Z_0, Z_1, \dots, Z_{n-1} be independent with the common exponential distribution of mean 1. For each $m \leq n$, set $S_m = Z_0 + \dots + Z_{m-1}$. Show that the normalised variables $Z_0/S_m, Z_1/S_m, \dots, Z_{m-1}/S_m$ are the spacings of $m-1$ random points in the unit interval $[0, 1]$.

11. *Continuation, a coupling.* Introduce notation $F(x) = P\{S_m \leq x\}$ and $G(x) = P\{S_m \leq x \mid S_m \leq s\}$ for the unconditional and conditional d.f.s of S_m , respectively. Write $H(S_m) = G^{-1}(F(S_m))$ for a random variable with the distribution G . Show that

$$\mathcal{L}(Z_0, Z_1, \dots, Z_{m-1} \mid S_m \leq s) = \mathcal{L}\left(H(S_m) \frac{Z_0}{S_m}, H(S_m) \frac{Z_1}{S_m}, \dots, H(S_m) \frac{Z_{m-1}}{S_m}\right).$$

12. *Continuation, the circular scan statistic.* Suppose $2 \leq m \leq n/2$ and $s > 0$. Let $X_j = 1(Z_j + Z_{j+1} + \dots + Z_{j+m-1} \leq s)$ where addition in the subscript is to be taken modulo n . The sum $W = X_0 + X_1 + \dots + X_{n-1}$ is called the circular scan statistic and has a variety of applications in medicine, signal processing, and spatial data analysis.¹⁸ By symmetry it suffices to construct a coupling for $\alpha = j = 0$. For $1 \leq k \leq m-1$ set

$$X'_{0k} = 1\left\{H(S_m)\left(\frac{Z_k}{S_m} + \frac{Z_{k+1}}{S_m} + \dots + \frac{Z_{m-1}}{S_m}\right) + Z_m + Z_{m+1} + \dots + Z_{m+k-1} \leq s\right\}.$$

Form similar couplings when $n-m \leq k \leq n-1$ and, finally, for $m \leq k \leq n-m-1$ set $X'_{0k} = X_k$. Using the result of the previous problem argue that the variables X_0, X_1, \dots, X_{n-1} are positively related, hence determine the accuracy of Poisson approximation.

For Problems 13 through 19 let $\xi_0, \xi_1, \dots, \xi_n$ denote random points on the circle \mathbb{T} of unit circumference and let L_1, \dots, L_{n+1} denote the corresponding spacings. Identify $n+1$ with 0.

13. *Spacings.* Verify the lemma at the end of Section 9 by showing that $P\{L'_2 > x_2, \dots, L'_{n+1} > x_{n+1}\} = P\{L_2 > x_2, \dots, L_{n+1} > x_{n+1} \mid L_1 > r\}$. [Hint: Evaluate the left-hand side by exploiting the independence of (M_2, \dots, M_{n+1}) and L'_1 .]

14. *Continuation, reconstructing the coupling.* Set $X'_{11} = X_1$ and $X'_{1k} = 1(L'_k > r)$ for $k \neq 1$ and deduce that $\mathcal{L}(X'_{12}, \dots, X'_{1,n+1}) = \mathcal{L}(X_2, \dots, X_{n+1} \mid X_1 = 1)$. Show that $L'_1 \geq L_1$ and hence that $L'_k \leq L_k$ for $k \neq 1$. Construct the entire coupling along these lines and argue that X_1, \dots, X_{n+1} are indeed negatively related.

15. *Isolation.* For each $0 \leq j \leq n$, form the indicator variables $X_j = 1(L_j > r, L_{j+1} > r)$. If $X_j = 1$ then the j th point is considered to be *isolated* from its neighbours. Show that the system of variables $X'_{jj} = X_j, X'_{j,j-1} = 1((1-2r)L_{j-1} > r), X'_{j,j+1} = 1((1-2r)L_{j+1} > r)$, and $X'_{jk} = 1((1-2r)L_j > r, (1-2r)L_{j+1} > r)$ for $k \notin \{j-1, j, j+1\}$ leads to a satisfactory coupling with $\Gamma_j^0 = \{j-1, j+1\}$ and $\Gamma_j^- = \{j+2, j+3, \dots, n+1, 1, 2, \dots, j-2\}$.

¹⁸See, for instance, S. E. Alm, "On the distribution of the scan statistic of a Poisson process", in *Probability and Mathematical Statistics: Essays in Honour of Carl-Gustav Esseen* (eds A. Gut and L. Holst), Department of Mathematics, Uppsala University, pp. 1-10, 1983.

16. Continuation, a threshold function for isolation. Find an estimate for the accuracy of Poisson approximation for the number of isolated points $W = X_0 + X_1 + \dots + X_{n+1}$ and thence a threshold function $r = r_n$ for the emergence of isolated points.

17. Small spacings. Introduce notation for the conditional d.f. $H(x) = P\{L_1 \leq x \mid L_1 \leq r\}$ and let $L''_1 = H^{-1}(F(L_1))$ denote a variable with the conditional distribution H . Form the variables $L''_k = (1 - L''_1)M_k$ for $k \neq 1$. Proceeding as for the lemma of Section 9 argue that L''_1 is independent of M_2, \dots, M_{n+1} and hence that $\mathcal{L}(L''_2, \dots, L''_{n+1}) = \mathcal{L}(L_2, \dots, L_{n+1} \mid L_1 \leq r)$.

18. Continuation, a new coupling. For $1 \leq j \leq n+1$, let $X_j = 1(L_j < r)$ be the indicator for a small spacing. Construct a coupling using the results of the previous problem and show that the variables X_1, \dots, X_{n+1} are negatively related. Thence determine the accuracy of Poisson approximation for $W = X_1 + \dots + X_{n+1}$, the number of small spacings.

19. Quarantining, complete isolation. Show that if $r = r_n = n^{-2}(e^{-c} + \eta_n)$ where $\eta_n \rightarrow 0$ then W has an asymptotic Poisson law with mean e^{-c} and *a fortiori* $P\{W = 0\} \rightarrow e^{-e^{-c}}$. If $W = 0$ there are no small spacings (as defined in the previous problem) which means that each of the points is completely isolated from the others in the sense that each point has a privacy buffer of at least $r = r_n$ on both sides. Invoking an epidemiological picture, if we think of the random points as potentially infective agents who can spread contagion in their immediate proximity over a distance of up to r then, for small r , as the number n of agents increases there is a threshold function manifest for n around the order $r^{-1/2}$ at which contagion is completely contained. Alternatively, this suggests a radius for quarantining a given population to contain disease.

The concluding problems of this chapter deal with an alternative, simple approach to a Poisson sieve which can be effective in some settings when the dependency structure is weak or sparse. The method constitutes a strengthening of the independence sieve of Section IV.9 and requires only correlative structures, obviating the need for an explicit coupling or the detailed calculation of conjunction probabilities. The structure is the following:

Suppose $\mathbf{X} = (X_1, \dots, X_n)$ is a random element in the space $\{0, 1\}^n$ endowed with product measure $p(\mathbf{x}) = P\{\mathbf{X} = \mathbf{x}\} = \prod_{j=1}^n p_j^{x_j} (1 - p_j)^{1-x_j}$. This induces a probability measure on subsets of $\{0, 1\}^n$ by the natural assignment $P(A) = \sum_{\mathbf{x} \in A} p(\mathbf{x})$ which sums the probabilities of all sequences in the given family A of sequences in $\{0, 1\}^n$. Monotone families of sequences will be of interest. As in Section XVII.4, say that a binary sequence \mathbf{y} is a *successor* of \mathbf{x} if $\mathbf{y} \geq \mathbf{x}$ and that \mathbf{y} is an *ancestor* of \mathbf{x} if $\mathbf{y} \leq \mathbf{x}$. Say that a family A of sequences in $\{0, 1\}^n$ is *increasing* if all successors of any sequence \mathbf{x} in A are also in A ; say that A is *decreasing* if all ancestors of any sequence \mathbf{x} in A are also in A . For any $\mathbf{x} \in \{0, 1\}^n$, write $S_\mathbf{x}$ for the smallest increasing family of sequences containing \mathbf{x} .

20. Kleitman's lemma. If A and B are either both increasing or both decreasing subsets of $\{0, 1\}^n$ then $P(A \cap B) \geq P(A)P(B)$. The inequality is reversed if one is increasing, the other decreasing.¹⁹ [Hint: Specialise Problem XIV.41 to indicator functions.]

¹⁹Kleitman's lemma was originally proved for uniform measures and was the starting point for the developments in correlation inequalities that followed: D. J. Kleitman, "Families of non-disjoint subsets", *Journal of Combinatorial Theory*, vol. 1, pp. 153–155, 1966.

21. *Janson's inequality.* Let S_{x_1}, \dots, S_{x_m} be successor families generated by given sequences $x_1, \dots, x_m \in \{0, 1\}^n$. For $1 \leq i \leq m$, let A_i denote the event that the random Bernoulli sequence $\mathbf{X} = (X_1, \dots, X_n)$ is contained in S_{x_i} . In the notation of Section IV.9, let J_i be the independence set of events A_j which are (jointly) independent of A_i . Write $M = \prod_i P(A_i^c)$, $\alpha = \max_i P(A_i)$, $\mu = \sum_i P(A_i)$, and $\Delta = \sum_i \sum_{j \notin J_i} P(A_i \cap A_j)$. Then

$$M \leq P(A_1^c \cap \dots \cap A_m^c) \leq M \exp\left(\frac{\Delta}{2(1-\alpha)}\right) \leq \exp\left(-\mu + \frac{\Delta}{2(1-\alpha)}\right).$$

Break up the proof of the two-sided inequality for $P(\bigcap_i A_i^c)$ into the following steps.²⁰
 (a) Using Kleitman's lemma argue that $P(A_i \mid \bigcap_{j \in J} A_j^c) \leq P(A_i)$ for all i and all subsets J and thence establish the lower bound by the chain rule for conditional probabilities. (b) For given i , renumber so that $J_i = \{d+1, d+2, \dots, n\}$ and write $B = \bigcap_{j \leq \min\{d, i-1\}} A_j^c$ and $C = \bigcap_{d+1 \leq j < i} A_j^c$. Justify the following steps:

$$P(A_i \mid B \cap C) \geq P(A_i \cap B \mid C) = P(B \mid A_i \cap C) P(A_i). \quad (\text{i})$$

$$P(B \mid A_i \cap C) \geq 1 - \sum_{j \leq d} P(A_j \mid A_i \cap C) \geq 1 - \sum_{j \leq d} P(A_j \mid A_i). \quad (\text{ii})$$

$$P\left(A_i \mid \bigcap_{j < i} A_j^c\right) \geq P(A_i) - \sum_{j \leq d} P(A_i \cap A_j). \quad (\text{iii})$$

$$P\left(A_i^c \mid \bigcap_{j < i} A_j^c\right) \leq P(A_i^c) \left(1 + \frac{1}{1-\alpha} \sum_{j \leq d} P(A_i \cap A_j)\right). \quad (\text{iv})$$

To finish up, use the chain rule again together with the elementary inequality $1+x \leq e^x$ to establish the upper bound. The factor $1/2$ in the exponent arises because in our definition of Δ each term $P(A_i \cap A_j)$ in the sum is counted twice.

22. *Continuation.* If, as $n \rightarrow \infty$, $\Delta \rightarrow 0$, $\alpha \rightarrow 0$, and $\mu \rightarrow e^{-c}$ for some constant c , then $P(A_1^c \cap \dots \cap A_m^c) \rightarrow e^{-e^{-c}}$ and we have asymptotic Poisson behaviour.

23. *Triangle-free graphs, once more.* Analyse the property that $G_{n,p}$ is triangle-free by Janson's inequality and marvel at the labour saved.

²⁰Janson discovered the inequality bearing his name during the *Third Conference on Random Graphs* (Poznań, July 1987); it appeared in S. Janson, T. Luczak, and A. Ruciński, "An exponential bound for the probability of nonexistence of a specified subgraph in a random graph", in *Random Graphs '87*, pp. 73–87, John Wiley, Chichester, 1990. Janson's original proof used Laplace transforms; the elementary method of proof suggested here is from R. Boppana and J. Spencer, "A useful elementary correlation inequality", *Journal of Combinatorial Theory A*, vol. 50, pp. 305–307, 1989.

XIX

Convergence in Law, Selection Theorems

We shall be concerned in this chapter primarily with sequences of distributions and their limiting properties. Our focus will be on distributions qua distributions, that is to say, on their *analytical* properties. The probabilistic origins of the distributions are not critical for these purposes but, of course, do much to add colour to the results and their character. There are, of course, also other notions of convergence that are important in probability—we've seen two examples in the weak law and the strong law of large numbers. Convergence in distribution (or *law*) is, in a certain sense, the weakest of these notions and it is perhaps not surprising hence that it comes equipped with a rich analytical structure that can be exploited.

C 1–3, 5, 6
F 4, 7–9

The material of this chapter is of a somewhat more abstract nature though the reader who has a sound grasp of what a continuous function is should have no difficulty in following along with the proofs. Scattered applications from a variety of domains are provided in the following chapter to help leaven the abstract fare. The rewards of abstraction at this stage are in the generality and power of the results that emerge and the wide range of application domains in which they hold sway. It comes as a pleasant surprise that the main results are susceptible of proof by elementary methods. To prepare the ground the reader should begin hence with a review of continuous functions; if she wishes to refresh her memory she will find the key facts summarised in Section XXI.2 in the Appendix.

I have intentionally left the more abstract material dealing with the subtle selection principle of Helly to the latter half of the chapter. The reader may elect to skim over this material on a first reading but she may well find that the beguiling applications of the principle merit more than a passing glance.

1 Vague convergence

The binomial distribution furnishes us with several examples of convergence of distributions. Suppose S_n denotes the number of successes in n tosses of a

coin with success probability p . As usual, with $q = 1 - p$, write $b_n(k; p) = \binom{n}{k} p^k q^{n-k}$ for the probability that S_n takes value k .

EXAMPLES: 1) *The weak law.* The scaled random variable S_n/n has atoms at the points k/n with corresponding probabilities $b_n(k; p)$. Let F_n denote the d.f. of S_n/n . In consequence of the weak law of large numbers, if \mathbb{I} is an interval and $n \rightarrow \infty$, then

$$F_n(\mathbb{I}) \rightarrow \begin{cases} 1 & \text{if } p \text{ is an interior point of } \mathbb{I}, \\ 0 & \text{if } p \text{ is an interior point of } \mathbb{I}^c, \end{cases}$$

as $n \rightarrow \infty$. (See Section V.6 and Problem V.17.) In words, F_n converges to the distribution concentrated at p over every interval for which p is not a boundary point.

2) *Poisson approximation.* Suppose λ is a positive constant and $p = p_n = \lambda/n$ is allowed to depend on n . Then for each fixed $k \geq 0$ we have $b_n(k; p) \rightarrow p(k; \lambda) = e^{-\lambda} \lambda^k / k!$ as $n \rightarrow \infty$ via the Poisson approximation to the binomial (Section VII.6). Indeed, by the theorem of Section XVIII.4, if we write F_n and F for the d.f.s of the binomial $b_n(k; p)$ and the Poisson $p(k; \lambda)$, respectively, and \mathbb{I} is any interval (or, indeed, any Borel set) on the line, then $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ for every interval \mathbb{I} . Thus, F_n converges to the Poisson distribution over every interval.

3) *The de Moivre–Laplace theorem.* The normalised random variable $S_n^* = (S_n - np)/\sqrt{npq}$ takes values only at the points $(k - np)/\sqrt{npq}$ with k varying over the integers from 0 through n , with the corresponding atomic probabilities $b_n(k; p)$ at these points. Reusing notation, let F_n now denote the d.f. of the variable S_n^* . Then de Moivre’s theorem of Section VI.5 (see also Problem VI.11) shows that $F_n(\mathbb{I}) \rightarrow \Phi(\mathbb{I})$ as $n \rightarrow \infty$ for every interval \mathbb{I} . Thus, F_n converges to the standard normal distribution over every interval.

4) *The arc sine law.* Suppose n is even for definiteness and in the notation of Section VIII.5 write $u_n = \binom{n}{n/2} 2^{-n}$ for the central term of the binomial. Consider the simple random walk $0 = S_0, S_1, \dots, S_n$ over n steps which starts at the origin and proceeds by taking a unit step left or right with equal probability at each epoch. Let M_n denote the index at which the walk achieves its maximum value for the first time and write F_n for the distribution of $\frac{1}{n} M_n$. In view of (VIII.5.5), we have

$$\mathbf{P}\left\{\frac{1}{n} M_n = \frac{k}{n}\right\} = \begin{cases} \frac{1}{2} u_k u_{n-k} & \text{if } k \text{ is even,} \\ \frac{1}{2} u_{k-1} u_{n-k+1} & \text{if } k \text{ is odd.} \end{cases}$$

By Theorem VIII.5.1, $F_n(t) = \mathbf{P}\left\{\frac{1}{n} M_n \leq t\right\} \rightarrow \frac{2}{\pi} \arcsin \sqrt{t}$ for each $0 < t < 1$ and so F_n converges to the arc sine distribution over every interval. ▶

What was important in these examples was not their coin-tossing genesis *per se* but the purely analytical properties of the distributions. In fact, the results could have been phrased solely in terms of the distributions without mention of the underlying probability spaces. (The random variables could even have existed in different probability spaces.) Casting such convergence statements in terms of distributions helps focus on the analytical ideas involved. In the following examples F denotes an arbitrary distribution function.

EXAMPLES: 5) Suppose $F_n(x) = F(x - n^{-1})$ for each n . Then $F_n(x) \rightarrow F(x)$ at each point of continuity of F while $F_n(x) \rightarrow F(x-)$ at points of jump of F . A little thought now shows that if \mathbb{I} is any interval whose endpoints are not points of jump of F then $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$.

6) If $F_n(x) = F(x + n)$ then $F_n(x) \rightarrow 1$ for all x . It follows that $F_n(\mathbb{I}) \rightarrow 0$ for every bounded interval. But, of course, $F_n(\mathbb{R}) = 1$. Here is an illustration where the d.f.s converge but *not to a limiting distribution function*. Another example can be furnished by setting $F_n(x) = F(x - n)$ whence $F_n(x) \rightarrow 0$ for all x . Again, $F_n(\mathbb{I}) \rightarrow 0$ for all bounded intervals but $F_n(\mathbb{R}) = 1$. Finally, suppose $F_n(x) = F(x + (-1)^n n)$. Then $F_n(x)$ does not converge for any x but $F_{2n}(x) \rightarrow 1$ and $F_{2n-1}(x) \rightarrow 0$ for all x . While the distribution functions as such do not converge, $F_n(\mathbb{I}) \rightarrow 0$ for every bounded interval while again $F_n(\mathbb{R}) = 1$. ►

Thus, a sequence of d.f.s $\{F_n\}$ may not converge, may converge but not to a d.f., or may converge to a limiting d.f. at points of continuity (but perhaps not at points of jump of the limiting d.f. in view).

Let F be any distribution function. For the rest of this chapter we will focus on F primarily in its rôle as a set function, that is to say, as a probability distribution or measure. We say that \mathbb{I} is an *interval of continuity* of F if \mathbb{I} is open and the endpoints of \mathbb{I} are not atoms, that is to say, points of jump of F . Thus, if $\mathbb{I} = (a, b)$ is an interval of continuity of F then $F(a, b) = F[a, b]$ and we can eschew the distinction between open, closed, and semi-closed intervals. For our purposes the entire line is also counted as an interval of continuity.

DEFINITION A sequence of probability distributions $\{F_n\}$ converges vaguely to a probability distribution F , denoted in brief by $F_n \xrightarrow{v} F$, if $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ for every bounded interval of continuity of F .¹ If the random variable X has d.f. F and, for each n , the random variable X_n has d.f. F_n , the statement X_n converges in distribution (or, *in law*) to X , written in brief as $X_n \xrightarrow{d} X$, is just another way of saying that $\{F_n\}$ converges vaguely to F .

If $\mathbb{I} = (a, b)$ is a bounded interval of continuity of F , then $F(\mathbb{I}) = F(b) - F(a)$ and it is clear that $F_n \xrightarrow{v} F$ if, and only if, $F_n(x) \rightarrow F(x)$ at each point of continuity of F . We could hence have chosen to express the idea as a property of

¹In higher dimensions all that is required is to replace intervals of continuity in the definition by bounded open rectangles whose boundary has probability zero with respect to F .

the point function instead. The choice in the definition reflects modern usage which tends to favour expressing the concept in terms of the more flexible set function instead of the classical point function. We shall generalise the setting very slightly in Section 6.

Note that if $F_n \xrightarrow{v} F$ then $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ for *unbounded* intervals of continuity of F as well. Indeed, as $F_n(-\infty, \infty) = F(-\infty, \infty) = 1$ there is nothing to be shown if \mathbb{I} is the entire line. Now suppose $\mathbb{I} = (a, \infty)$ is a half-line. We can select $b > |a|$ and an interval of continuity $(-b, b)$ of F so large that $F(-b, b) > 1 - \epsilon$. But $|F_n(-b, b) - F(-b, b)| < \epsilon$ for all sufficiently large n as $F_n \xrightarrow{v} F$. In particular, $F_n[b, \infty) < 2\epsilon$. It follows that $|F_n(a, \infty) - F(a, \infty)| \leq |F_n(a, b)| + F_n[b, \infty) + F[b, \infty) < 4\epsilon$ and as ϵ is arbitrary, $F_n(a, \infty) \rightarrow F(a, \infty)$. We can handle unbounded intervals of the form $(-\infty, b)$ in the same way. Thus, if $F_n \xrightarrow{v} F$ then $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ for all intervals of continuity of F , bounded or unbounded.

2 An equivalence theorem

The abstract notion of distributional convergence is made much more concrete by relating it to expectations of continuous functions of random variables. As preparation, we begin with the family \mathcal{C} of continuous, real-valued functions of a real variable. This class is too rich and uncontrolled for our purposes and we successively whittle it down to a function family rich enough to be useful and regular enough to be analytically tractable. We write \mathcal{C}_b for the subfamily of continuous, bounded, real-valued functions of a real variable. This class still can have too much variability at infinity and accordingly we will restrict attention further to the family $\bar{\mathcal{C}}_b$ of continuous functions u with finite limits at infinity, that is to say, the limits $u(-\infty)$ and $u(+\infty)$ (exist and) are finite. If, in addition, $u(-\infty) = u(+\infty) = 0$ then we say that the function u vanishes at infinity. We will focus on a subclass of $\bar{\mathcal{C}}_b$ that is even more analytically malleable. We say that a function u is *smooth* if it is infinitely differentiable and the function and each of its derivatives has finite limits at infinity. This last is the class we seek.

In the preceding we have asserted that the classes of functions that we have introduced are nested. Only the assertion $\bar{\mathcal{C}}_b \subset \mathcal{C}_b$ requires any justification and we will actually prove a little more than claimed: *the family $\bar{\mathcal{C}}_b$ is a uniformly continuous subclass of the family of bounded, continuous functions \mathcal{C}_b .*² Indeed, suppose $u \in \bar{\mathcal{C}}_b$. We may suppose that $u \geq 0$ (else consider $|u|$ instead). Then, u has the finite limits $\lim_{x \rightarrow \infty} u(x) = M_\infty$ and $\lim_{x \rightarrow -\infty} u(x) = M_{-\infty}$. In particular, for any $\epsilon > 0$, there exists a such that $|u(x) - M_\infty| < \epsilon$ for

²The real line is said to be *compactified* if the points $\pm\infty$ are appended to it. We may note in brief that $\bar{\mathcal{C}}_b$ is identically the family of continuous functions on the compactified line. In v dimensions, the family $\bar{\mathcal{C}}_b(v)$ may be specified recursively as the class of continuous functions $u(x) = u(x_1, \dots, x_v)$ for which the limits as $x_k \rightarrow \pm\infty$ exist and are functions in the class $\bar{\mathcal{C}}_b(v-1)$ in one lower dimension.

$x > a$ and $|u(x) - M_{-\infty}| < \epsilon$ for $x < -a$. As u is continuous it achieves its maximum, say M_a , over the closed interval $[-a, a]$. Consequently, $\sup u(x) \leq \max\{M_{-\infty} + \epsilon, M_a, M_\infty + \epsilon\}$ is finite and u is bounded. To establish that u is uniformly continuous, we observe that $|u(x) - u(y)| < 2\epsilon$ if x and y are both larger than a or both less than $-a$. It follows that u is uniformly continuous on $(-\infty, a) \cup (a, \infty)$. As the interval $[-a, a]$ is closed and bounded, u is uniformly continuous in this interval, as also in the intervals $[-a-r, -a+r]$ and $[a-r, a+r]$ for any positive r . It follows easily that u is bounded and uniformly continuous on the entire line.

A key result links vague convergence to convergence of a sequence of expectations for the family of functions \mathcal{C}_b . Here and elsewhere, unless specified otherwise integrals are to be taken over the entire space.

EQUIVALENCE THEOREM I *Let $\{F_n\}$ and F be distributions. Then $F_n \xrightarrow{v} F$ if, and only if, for every $u \in \mathcal{C}_b$,*

$$\int u(x) dF_n(x) \rightarrow \int u(x) dF(x). \quad (2.1)$$

If random variables $\{X_n\}$ and X correspond to the distributions $\{F_n\}$ and F , respectively, then (2.1) is equivalent to saying that $E(u(X_n)) \rightarrow E(u(X))$.

PROOF: To prove sufficiency, suppose $\mathbb{I} = (a, b)$ is a bounded interval of continuity of F . We wish to show that $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ assuming (2.1) holds for all $u \in \mathcal{C}_b$. Suppose now that $u = 1_{(a,b)}$ is the indicator for \mathbb{I} . Then the statement $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ is equivalent to saying that $\int u dF_n \rightarrow \int u dF$. We can't directly appeal to (2.1) here as our u is not continuous. P. Lévy's ingenious idea to get around this difficulty was to sandwich u between two continuous functions and then squeeze u by making the sandwich thinner and thinner.

As (a, b) is an interval of continuity of F , for a sufficiently small, positive h we have $F(a-h, a+h) < \epsilon$ and $F(b-h, b+h) < \epsilon$. Now consider the continuous functions u_+ and u_- which between them sandwich the indicator u as shown in Figure 1. The reader has seen this pretty idea put to use in Section VI.5 and it continues to work in this setting for much the same reasons.

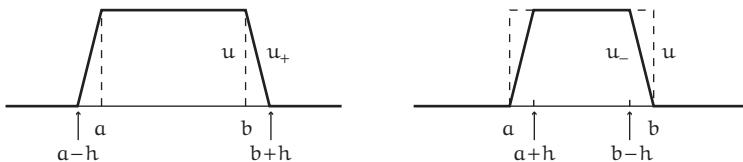


Figure 1: Lévy sandwich.

It is clear that $u_- \leq u \leq u_+$ whence by monotonicity of expectation, $\int u_- dF_n \leq \int u dF_n \leq \int u_+ dF_n$ and $\int u_- dF \leq \int u dF \leq \int u_+ dF$. It is of course understood that the integrals, unless specified otherwise, are over the entire real line \mathbb{R} . As u_- and u_+ are continuous and bounded, we have $\int u_- dF_n \rightarrow \int u_- dF$ and $\int u_+ dF_n \rightarrow \int u_+ dF$. In particular, $\int u_- dF_n > \int u_- dF - \epsilon$ and $\int u_+ dF_n < \int u_+ dF + \epsilon$ for all sufficiently large n . We take a little notational liberty and combine two strings of inequalities into one to write

$$0 \leq \int |u_\pm - u| dF \leq F(a-h, a+h) + F(b-h, b+h) < 2\epsilon.$$

And so, for all sufficiently large n ,

$$\int u dF - 3\epsilon < \int u_- dF_n \leq \int u dF_n \leq \int u_+ dF_n < \int u dF + 3\epsilon.$$

Thus, $|\int u dF_n - \int u dF| < 3\epsilon$, eventually. As $\epsilon > 0$ is arbitrary, it follows that $F_n(\mathbb{I}) = \int u dF_n \rightarrow \int u dF = F(\mathbb{I})$.

In order to prove necessity, suppose $F_n \xrightarrow{v} F$. We may select a bounded interval of continuity $\mathbb{I} = (a, b)$, say, of F so large that $F(\mathbb{I}) > 1 - \epsilon$. As $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$, it follows that $F_n(\mathbb{I}) > 1 - 2\epsilon$, eventually. Now suppose u is any member of the class \mathcal{C}_b . Then, by the triangle inequality,

$$\left| \int u dF_n - \int u dF \right| \leq \left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} u dF \right| + \int_{\mathbb{I}^c} |u| dF_n + \int_{\mathbb{I}^c} |u| dF.$$

As $u \leq M$ for some M , $\int_{\mathbb{I}^c} |u| dF_n \leq M F_n(\mathbb{I}^c) < 2M\epsilon$ and, likewise, $\int_{\mathbb{I}^c} |u| dF \leq M F(\mathbb{I}^c) < M\epsilon$. It follows that

$$\left| \int u dF_n - \int u dF \right| \leq \left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} u dF \right| + 3M\epsilon.$$

As u is continuous, its restriction to the closed and bounded interval $[a, b]$ is uniformly continuous. Accordingly, we may partition \mathbb{I} into a finite number of subintervals $\mathbb{I}_1, \dots, \mathbb{I}_m$ so that the oscillation of u in any subinterval \mathbb{I}_k is less than ϵ . We may suppose without loss of generality that each \mathbb{I}_k is an interval of continuity of F and we consider n large enough that $|F_n(\mathbb{I}_k) - F(\mathbb{I}_k)| < \epsilon/m$ for each k . Now let $s = \sum_{k=1}^m u(x_k) 1_{\mathbb{I}_k}$ be any simple function where $x_k \in \mathbb{I}_k$ for each k . By the triangle inequality,

$$\left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} u dF \right| \leq \left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} s dF_n \right| + \left| \int_{\mathbb{I}} s dF_n - \int_{\mathbb{I}} s dF \right| + \left| \int_{\mathbb{I}} s dF - \int_{\mathbb{I}} u dF \right|. \quad (2.2)$$

We may partition the domain of integration \mathbb{I} into the subintervals \mathbb{I}_k to successively estimate each of the terms on the right. For the middle term, we have

$$\int_{\mathbb{I}} s dF_n - \int_{\mathbb{I}} s dF = \sum_{k=1}^m u(x_k) (F_n(\mathbb{I}_k) - F(\mathbb{I}_k))$$

by the definition of expectation of simple functions. It follows that $|\int_{\mathbb{I}} s dF_n - \int_{\mathbb{I}} s dF| < M\epsilon$. As for the remaining two terms, by additivity, the first term on the right of (2.2) is bounded by

$$\begin{aligned} \left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} s dF_n \right| &= \left| \sum_{k=1}^m \int_{\mathbb{I}_k} (u - u(x_k)) dF_n \right| \\ &\leq \sum_{k=1}^m \int_{\mathbb{I}_k} |u - u(x_k)| dF_n < \epsilon \sum_{k=1}^m F_n(\mathbb{I}_k) = \epsilon F_n(\mathbb{I}) < \epsilon, \end{aligned}$$

and an entirely similar argument shows that the third term on the right of (2.2) may be bounded by $|\int_{\mathbb{I}} s dF - \int_{\mathbb{I}} u dF| < \epsilon$. Combining our estimates we obtain the bound $|\int u dF_n - \int u dF| < (4M + 2)\epsilon$ and as ϵ is arbitrary, the upper bound can be made as small as desired. This establishes (2.1). ▶

If F and $\{F_n\}$ are distributions in v dimensions then a cursory examination shows that the proof of the theorem goes through provided we identify \mathcal{C}_b with the family of bounded, continuous functions in v dimensions. All that is required is to strike out sundry references to intervals in the proof and replace them with rectangles in v dimensions, and to replace integrals over \mathbb{R} by integrals over \mathbb{R}^v . For clarity of presentation I will continue to deal with the one-dimensional case; the reader should bear in mind, however, that the results go through generally to higher dimensions with obvious adaptations in terminology and notation.

Our theorem hence suggests a more flexible way of defining vague convergence in one or more dimensions and some nomenclature has built up in this regard. If (2.1) holds for all $u \in \mathcal{C}_b$ then this is sometimes expressed by saying that F_n converges weakly to F with respect to the class \mathcal{C}_b . Thus, vague convergence is equivalent to weak convergence with respect to a suitable function class.³

3 Convolutional operators

The convolution relation (XIV.10.2) for distributions suggests that it may be profitable to consider more general convolutions by replacing the distribution function in the integrand by an (integrable) function. This is the starting point of a rich and profitable development.

Suppose F is any probability distribution. If u is an integrable function (with respect to F) then $E(u) = \int u dF$. To ensure integrability it will suffice for us to consider u in the class \mathcal{C}_b of bounded, continuous functions. For each real t , we now define the function $u_t(x) = u(t - x)$ related to u by a reflection of coordinate and shift of origin. Then u_t is in \mathcal{C}_b , hence integrable, and $E(u_t) =$

³For a comprehensive exploration of these ideas see P. Billingsley, *Convergence of Probability Measures*, Second Edition. New York: John Wiley, 1999.

$\int u_t dF = \int u(t-x) dF(x)$ expresses a generalised convolution integral. The convolution notation, if used, then emends to $F * u$ for the expression on the right. The extension of the convolution notation to this setting is not entirely satisfactory—aside from being slightly cumbrous, the expression $F * u$ is not commutative—and it is preferable in this context to introduce a new notation and terminology.

As t varies from $-\infty$ to $+\infty$, the expectations $\int u_t dF$ determine a function of the variable t . As the nature of this function is determined by the distribution F , to each probability distribution F we may hence associate the *convolutional operator* $\mathfrak{F} : u \mapsto \mathfrak{F}u$ on the space \mathcal{C}_b of bounded, continuous functions which maps each function $u \in \mathcal{C}_b$ into another function $\mathfrak{F}u$ defined by

$$(\mathfrak{F}u)(t) = \int_{-\infty}^{\infty} u(t-x) dF(x)$$

for each t . The reader should immediately observe that, as the convolutional operator \mathfrak{F} merely implements expectation with respect to the distribution F , additivity of expectation immediately implies that *each convolutional operator \mathfrak{F} is a linear map on the space of bounded, continuous functions*. In particular, if u and v are functions in the class \mathcal{C}_b and λ is any real scalar, then $\mathfrak{F}(u+v) = \mathfrak{F}u + \mathfrak{F}v$ and $\mathfrak{F}(\lambda u) = \lambda(\mathfrak{F}u)$.

Insofar as is possible, as a notational *convention*, we will use capital Latin letters for distributions and corresponding capital Gothic letters for the associated convolutional operators; thus, we associate the distribution-operator pairs, (F, \mathfrak{F}) , (G, \mathfrak{G}) , (H, \mathfrak{H}) , etc. These notations and conventions were introduced by W. Feller.

The equivalence theorem of the previous section can now be restated to provide a slightly more malleable equivalent condition for vague convergence in terms of the associated convolutional operators on the better-behaved class $\bar{\mathcal{C}}_b$ of functions that are continuous, bounded, and have finite limits at $\pm\infty$.

EQUIVALENCE THEOREM II *Suppose $\{F_n\}$ is a sequence of probability distributions with corresponding operators $\{\mathfrak{F}_n\}$, F another probability distribution with corresponding operator \mathfrak{F} . Then $F_n \xrightarrow{v} F$ if, and only if, $\mathfrak{F}_n u \rightarrow \mathfrak{F}u$ uniformly for every $u \in \bar{\mathcal{C}}_b$.*

PROOF: It is only needful to retrace the steps in the proof of the equivalence theorem. For sufficiency, observe that if we assume, for any $u \in \bar{\mathcal{C}}_b$, that $\mathfrak{F}_n u \rightarrow \mathfrak{F}u$ uniformly then *a fortiori* $\int u dF_n \rightarrow \int u dF$. (If the continuous function u has finite limits at $\pm\infty$ then so does the continuous function $\hat{u}(x) = u(-x)$; as $(\mathfrak{F}_n \hat{u})(t) \rightarrow (\mathfrak{F}\hat{u})(t)$ uniformly in t , our conclusion follows by setting $t = 0$.) In particular, we may select as u the sandwich functions u_- and u_+ in the proof of the equivalence theorem and the rest of the proof remains unaltered.

In the proof of necessity in the equivalence theorem we were required to partition the interval of continuity \mathbb{I} of F into finite subintervals of continuity $\mathbb{I}_1, \dots, \mathbb{I}_m$ in such a way that the oscillation of u was less than ϵ in each

subinterval \mathbb{I}_k . Observe now that the functions $u_t(x) = u(t - x)$ are uniformly continuous for every $u \in \bar{\mathcal{C}}_b$ per our observation above and, indeed, the family of functions $\{u_t, t \in \mathbb{R}\}$ obtained by allowing the parameter t to vary over all real values is *equicontinuous*. It follows that we may select the subintervals of continuity \mathbb{I}_k so that the largest oscillation of any u_t is less than ϵ in any of the subintervals \mathbb{I}_k , in notation, $\max_{1 \leq k \leq m} \text{osc}_{\mathbb{I}_k} u_t < \epsilon$ for every t . (This is the reason for restricting attention to the family $\bar{\mathcal{C}}_b$ instead of the bigger family \mathcal{C}_b : we require uniform continuity.) The rest of the proof proceeds as before. ►

The idea of a function norm helps simplify notation and clarify concepts. Suppose u is any function. We define the (*supremum*) *norm* of u , written $\|u\|$, by $\|u\| := \sup_x |u(x)|$, with the usual convention that we write $\|u\| = \infty$ if the function is unbounded. It is easy to see that the norm satisfies the usual properties of length. Indeed, if u and v are any two functions and λ any real scalar, then the norm satisfies the properties of *positivity*, $\|u\| \geq 0$ with equality if, and only if, $u = 0$, the *scaling law*, $\|\lambda u\| = |\lambda| \cdot \|u\|$, and the *triangle inequality*, $\|u + v\| \leq \|u\| + \|v\|$. The only assertion that needs any justification is the triangle inequality and this follows quickly from the observation that $|u(x) + v(x)| \leq |u(x)| + |v(x)|$ for every x whence $\sup_x |u(x) + v(x)| \leq \sup_x (|u(x)| + |v(x)|) \leq \sup_x |u(x)| + \sup_x |v(x)|$.

Uniform convergence is succinctly described in terms of norms: *a sequence of functions $\{u_n\}$ converges uniformly to the function u if, and only if, $\|u_n - u\| \rightarrow 0$* . Now suppose \mathfrak{F}_n and \mathfrak{F} are operators corresponding to the distributions F_n and F . We write $\mathfrak{F}_n \rightarrow \mathfrak{F}$ to mean that $\|\mathfrak{F}_n u - \mathfrak{F} u\| \rightarrow 0$ for every $u \in \bar{\mathcal{C}}_b$; of course, this is equivalent to saying that $\mathfrak{F}_n u \rightarrow \mathfrak{F} u$ uniformly for every $u \in \bar{\mathcal{C}}_b$. With this notation, the equivalence theorem may be compactly expressed as follows: $F_n \xrightarrow{v} F$ if, and only if, $\mathfrak{F}_n \rightarrow \mathfrak{F}$.

An approximation principle of considerable power arises out of the theorem. Let H_0 denote the Heaviside distribution concentrated at the origin, \mathfrak{H}_0 its associated operator. As $H_0(x)$ has a single jump of size 1 at the origin, we have $\mathfrak{H}_0 u(t) = \int u(t - x) dH_0(x) = u(t)$, or in short, $\mathfrak{H}_0 u = u$, for every continuous u . Suppose F is any distribution. For each $h > 0$, we form the distribution F_h defined by $F_h(x) = F(x/h)$. It is clear that F_h converges vaguely to H_0 as $h \rightarrow 0$. In consequence of the equivalence theorem, this is the same as saying that $\mathfrak{F}_h \rightarrow \mathfrak{H}_0$. Or, explicitly, for every $u \in \bar{\mathcal{C}}_b$,

$$\mathfrak{F}_h u(t) = \int_{-\infty}^{\infty} u(t - x) dF_h(x) \rightarrow u(t) \quad (h \rightarrow 0), \quad (3.1)$$

the convergence being uniform in t . If F is absolutely continuous with density f this is the same as saying

$$\mathfrak{F}_h u(t) = \int_{-\infty}^{\infty} u(t - x) \frac{1}{h} f\left(\frac{x}{h}\right) dx = \int_{-\infty}^{\infty} u(x) \frac{1}{h} f\left(\frac{t-x}{h}\right) dx \rightarrow u(t)$$

uniformly in t . If we select for F the standard normal distribution Φ then its density $f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ has bounded derivatives of all orders, each of which is absolutely integrable and vanishes at infinity. It follows by Theorem XXI.2.4 in the Appendix that we may differentiate as often as we wish under the integral sign so that $\mathfrak{F}_h u$ also has derivatives of all orders, each derivative possessed of bounded limits at infinity.

Recasting our conclusions in this terminology, we obtain a uniform approximation theorem as an unexpected, and useful, bonus.

CONVOLUTIONAL SMOOTHING THEOREM *For any fixed $u \in \bar{\mathcal{C}}_b$ and any $\epsilon > 0$, there exists a smooth function v such that $\|u - v\| < \epsilon$.*

Thus, any $u \in \bar{\mathcal{C}}_b$ can be uniformly approximated by an infinitely differentiable function. Our analysis actually goes a little further and identifies a family of infinitely differentiable approximating functions $\mathfrak{F}_h u$ whose derivatives are all members of the class $\bar{\mathcal{C}}_b$ and which are defined via the convolution

$$\mathfrak{F}_h u(t) = \frac{1}{h\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x) \exp\left(\frac{-(t-x)^2}{2h^2}\right) dx.$$

The functions $\mathfrak{F}_h u$ given above are said to be obtained from u by smoothing with respect to a *Gaussian kernel*. Our uniform approximation theorem says that $\mathfrak{F}_h u$ converges uniformly to u as $h \rightarrow 0$. The reader who has read Section XVI.5 will have seen approximations of a rather different character.



4 An inversion theorem for characteristic functions

The approximation principle embodied in (3.1) has other deep ramifications. We consider in this section a fundamental uniqueness theorem which would otherwise require considerable effort to prove.

The reader has seen in Chapter XV that the Laplace transform of a distribution can be a very useful tool in divers applications but is sadly limited to settings with positive variables. Moving to the Fourier transform eliminates this annoying limitation. We write $i = \sqrt{-1}$ as usual and reuse the $\hat{\cdot}$ notation for the transform, there being no danger of confusion as long as we bear in mind what hat (sic) the function is wearing.

DEFINITION The *characteristic function* (c.f.) of a random variable X with (arbitrary) d.f. F is defined to be the complex-valued function $\hat{F}(\xi)$ of a real variable ξ given by

$$\hat{F}(\xi) = \mathbb{E}(e^{i\xi X}) = \int_{-\infty}^{\infty} e^{i\xi x} dF(x). \quad (4.1)$$

If F is absolutely continuous with density f then \hat{F} is just the ordinary Fourier transform of f introduced in Section VI.2 but we won't muddy the notational waters any further and keep to our slightly overburdened notation \hat{F} for the characteristic function.

The Fourier transform inherits most of the properties familiar from the Laplace transform (see Problems XV.28–30) but has the great advantage of not being limited to positive variables: as $e^{i\xi x}$ has unit modulus, the integral on the right in (4.1) is well-defined for all ξ . The price we pay is in the correspondingly heavy machinery of complex variable theory.

While (4.1) provides a map $F \mapsto \widehat{F}$, it is not clear whether it is one-to-one, that is to say, whether, to each d.f. F is associated a *unique* Fourier transform \widehat{F} . If this were the case then the relation (4.1) could, in principle, be inverted to recover F from \widehat{F} . The reader conversant with Fourier theory will recall that proving uniqueness is not at all a trivial matter; the fact that we are dealing with the transform of *distributions* makes all the difference.

We may write (4.1) in the form $e^{-i\xi t} \widehat{F}(\xi) = \int_{-\infty}^{\infty} e^{i\xi(x-t)} dF(x)$ by multiplying both sides by $e^{-i\xi t}$. Now suppose G is any distribution. If we treat ξ as a random variable with d.f. G , by integrating out with respect to G , we obtain

$$\begin{aligned} \int_{-\infty}^{\infty} e^{-i\xi t} \widehat{F}(\xi) dG(\xi) &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} e^{i\xi(x-t)} dF(x) \right\} dG(\xi) \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} e^{i\xi(x-t)} dG(\xi) \right\} dF(x) = \int_{-\infty}^{\infty} \widehat{G}(x-t) dF(x). \end{aligned}$$

The interchange of integrals is easily seen to be permitted here by Fubini's theorem, again as $e^{i\xi(x-t)}$ has modulus one. The integral identity takes on a particularly useful form if we select for G the normal distribution with mean 0 and variance $1/h^2$. Here we think of h as a small positive quantity so that G is very spread out. Then $G(\xi) = \Phi(h\xi)$, $dG(\xi) = h\phi(h\xi) d\xi$ and, in view of Example VI.2.4, $\widehat{G}(x-t) = e^{-(x-t)^2/2h^2} = \sqrt{2\pi} \phi\left(\frac{x-t}{h}\right)$. It follows that, for every $h > 0$, we have

$$\int_{-\infty}^{\infty} e^{-i\xi t} \widehat{F}(\xi) h\phi(h\xi) d\xi = \sqrt{2\pi} \int_{-\infty}^{\infty} \phi\left(\frac{x-t}{h}\right) dF(x)$$

or, what is the same thing,

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\xi t} \widehat{F}(\xi) e^{-h^2 \xi^2/2} d\xi = \int_{-\infty}^{\infty} \frac{1}{h} \phi\left(\frac{t-x}{h}\right) dF(x). \quad (4.2)$$

(We may replace $x-t$ in the argument of ϕ on the right by $t-x$ as ϕ is even.) Introducing the nonce notation Φ_h for the normal distribution with mean zero and variance h^2 , by the fundamental theorem of calculus, we may identify the expression on the right with

$$\begin{aligned} \frac{d}{dt} \int_{-\infty}^t \left\{ \int_{-\infty}^{\infty} \frac{1}{h} \phi\left(\frac{\tau-x}{h}\right) dF(x) \right\} d\tau &= \frac{d}{dt} \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^t \frac{1}{h} \phi\left(\frac{\tau-x}{h}\right) d\tau \right\} dF(x) \\ &= \frac{d}{dt} \int_{-\infty}^{\infty} \Phi\left(\frac{t-x}{h}\right) dF(x) = \frac{d}{dt} (F * \Phi_h)(t), \end{aligned}$$

the change in order of integration again easily justified by Fubini's theorem. It follows that the expression on the right in (4.2) may be identified with the density f_h of the distribution F_h obtained as the convolution of F with Φ_h . By integration of the density we

may recover $F_h = F * \Phi_h$ which, as convolution of distributions commutes, is the same as the distribution $\Phi_h * F$. We have hence shown that for every $h > 0$, the characteristic function \widehat{F} uniquely determines the distribution $F_h = \Phi_h * F$ through the relation (4.2). But Φ_h has variance h^2 and so, as $h \rightarrow 0$, Φ_h converges vaguely to the Heaviside distribution H_0 . We hence have $F_h(t) \rightarrow F(t)$ as $h \rightarrow 0$, at least at points of continuity of F . (This is merely a variant of our basic approximation principle (3.1); if the reader is sceptical she is invited to partition the region of integration into $|x| < \delta$ and $|x| \geq \delta$ and, proceeding in the usual way, argue that the contribution from $|x| < \delta$ is small because of the continuity of F at t and the contribution from $|x| \geq \delta$ is small because of the rapid decay of the normal tail.) The right continuity of F specifies the values of F then at points of jump as well. It follows that the characteristic function \widehat{F} uniquely determines the distribution F .

INVERSION THEOREM FOR CHARACTERISTIC FUNCTIONS *Every d.f. F has a unique characteristic function \widehat{F} .*

Of course, the reader would like an explicit inversion formula and (4.2) contains the germ of one and more besides.

COROLLARY *If \widehat{F} is integrable then F has a continuous, bounded density f given by the inversion formula*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{F}(\xi) e^{-i\xi x} d\xi. \quad (4.3)$$

PROOF: Write $f_h(t)$ for the density appearing on the right of (4.2). Integration gives the corresponding distribution $F_h(\mathbb{I}) = \int_{\mathbb{I}} f_h(t) dt = \Phi_h * F(\mathbb{I})$ for every interval (or even Borel set) \mathbb{I} . As $F_h \xrightarrow{v} F$ we have $F_h(\mathbb{I}) \rightarrow F(\mathbb{I})$ for every bounded interval of continuity of F . On the other hand, if \widehat{F} is integrable, then, by allowing h to tend to zero on the left-hand side of (4.2), we see that f_h converges boundedly and uniformly to the continuous function f given by (4.3). (The uniform limit of continuous functions is continuous.) Thus, for every bounded interval \mathbb{I} , we have $\int_{\mathbb{I}} f_h(t) dt \rightarrow \int_{\mathbb{I}} f(t) dt$ and we conclude that the d.f. F has the bounded, continuous density f given by (4.3). ▶

A wide variety of other inversion formulæ may be crafted but have limited utility.

5 Vector spaces, semigroups

In general, an operator \mathfrak{T} on the space \mathcal{C}_b of bounded, continuous functions is a map which takes each $u \in \mathcal{C}_b$ into a function $\mathfrak{T}u$. A generic operator \mathfrak{T} need not correspond to a probability distribution, that is to say, need not be expressible as a convolution, and, in general, will not be linear.

THE VECTOR SPACE OF BOUNDED OPERATORS

We say that an operator $\mathfrak{T}: u \mapsto \mathfrak{T}u$ is *bounded* if there exists a positive constant τ such that $\|\mathfrak{T}u\| \leq \tau \|u\|$ for all $u \in \mathcal{C}_b$, where, as before, $\|u\|$ and $\|\mathfrak{T}u\|$ denote the (supremum) norms of the functions u and $\mathfrak{T}u$, respectively. We may think

of τ as a bound on the maximum “magnification” $\|\mathfrak{T}u\|/\|u\|$ for any non-zero u . In particular, if $\|u\| \leq M$ then $\|\mathfrak{T}u\| \leq \tau M$ and the image under a bounded operator \mathfrak{T} of a bounded, continuous function u is a bounded function $\mathfrak{T}u$.

As an immediate consequence of the definition, it is clear that if \mathfrak{T} is bounded then it takes the function $u_0(t) = 0$ that is zero everywhere back into itself, $\mathfrak{T}: u_0 \mapsto u_0$.

If \mathfrak{S} and \mathfrak{T} are two bounded operators, we may define operator *addition* by the natural assignment $(\mathfrak{S} + \mathfrak{T}): u \mapsto \mathfrak{S}u + \mathfrak{T}u$. As it is clear that the operator sum of two bounded operators is bounded it follows that the family of bounded operators is closed under addition. The reader will readily verify that operator addition has the properties we expect of addition—it is commutative, $\mathfrak{S} + \mathfrak{T} = \mathfrak{T} + \mathfrak{S}$, associative, $\mathfrak{R} + (\mathfrak{S} + \mathfrak{T}) = (\mathfrak{R} + \mathfrak{S}) + \mathfrak{T}$, and has as zero element the zero operator $\mathfrak{O}: u \mapsto u_0$ which maps each function u to $u_0 = 0$.

The scalar multiplication of a bounded operator \mathfrak{T} is, likewise, defined naturally by the assignment $\lambda\mathfrak{T}: u \mapsto \lambda(\mathfrak{T}u)$ for every real scalar λ . It is again manifest that the family of bounded operators is closed under scalar multiplication.

As usual, we identify the operator difference $\mathfrak{S} - \mathfrak{T}$ with the operator $\mathfrak{S} + (-1)\mathfrak{T}$. While we will be mainly concerned with sums and differences of convolutional operators, it is clear that a much wider range of operators can be constructed by repeated applications of sums and scalar products.

We may hence freely add, subtract, and scale bounded operators just as we would ordinary vectors. It follows that the family of bounded operators \mathfrak{T} on the class \mathcal{C}_b of bounded, continuous functions forms a linear vector space. It is now natural to ask whether we can define a notion of length on this space.

DEFINITION The *norm* of a bounded operator \mathfrak{T} on the class \mathcal{C}_b of bounded, continuous functions is denoted $\|\mathfrak{T}\|$ and defined by

$$\|\mathfrak{T}\| := \inf\{\tau : \|\mathfrak{T}u\| \leq \tau\|u\|, u \in \mathcal{C}_b\}.$$

It is simple to verify that the basic properties of length are satisfied. Let \mathfrak{S} and \mathfrak{T} be bounded operators, λ a real constant.

1. *Positivity*: $\|\mathfrak{T}\| \geq 0$ with equality if, and only if, \mathfrak{T} is the zero operator \mathfrak{O} .
2. *Scaling law*: $\|\lambda\mathfrak{T}\| = |\lambda| \cdot \|\mathfrak{T}\|$.
3. *Triangle inequality*: $\|\mathfrak{S} + \mathfrak{T}\| \leq \|\mathfrak{S}\| + \|\mathfrak{T}\|$.

As an immediate consequence of the definition, it follows that $\|\mathfrak{T}u\| \leq \|\mathfrak{T}\| \cdot \|u\|$ for every u . The norms at either end are the supremum norms of the associated functions, while the norm in the middle is the operator norm. It is unnecessary to introduce new notation to differentiate between the function and operator norms; the argument makes clear which norm is under consideration.

Let us now return to convolutional operators. Suppose, in our usual notation, that \mathfrak{F} is the operator corresponding to a probability distribution F . Then, by the modulus inequality for expectations,

$$|\mathfrak{F}u(t)| \leq \int_{-\infty}^{\infty} |u(t-x)| dF(x) \leq \|u\| \int_{-\infty}^{\infty} dF(x) = 1 \cdot \|u\|$$

for every $u \in \mathcal{C}_b$. In consequence, $\|\mathfrak{F}\| \leq 1$ and \mathfrak{F} is a bounded operator. On the other hand, if u takes a constant value M at all points, then $\|\mathfrak{F}u\| = M = 1 \cdot \|u\|$ and it follows that $\|\mathfrak{F}\| \geq 1$. It must be the case therefore that $\|\mathfrak{F}\| = 1$ identically and consequently *the operators corresponding to distributions all have norm 1*.

When dealing with sums and scalar multiples of operators, the reader should bear in mind that the operators $\mathfrak{F} + \mathfrak{G}$ and $\lambda\mathfrak{F}$ are not in general associated with probability distributions (even if \mathfrak{F} and \mathfrak{G} are). She will, however, be readily able to verify that, for every $0 \leq \lambda \leq 1$, the convex combination $\lambda\mathfrak{F} + (1 - \lambda)\mathfrak{G}$ of two convolutional operators \mathfrak{F} and \mathfrak{G} is the convolutional operator corresponding to the mixture distribution $\lambda F + (1 - \lambda)G$.

We conclude this section with an illustration of the ease of manoeuvring with the operator notation. The reader will recall that a smooth function v is infinitely differentiable with v and each of its derivatives $v^{(n)}$ possessed of finite limits at infinity. The fact that, *vide* the convolutional smoothing theorem of the previous section, any function u in the subclass $\bar{\mathcal{C}}_b$ of continuous functions with limits at infinity can be uniformly approximated by a smooth function v can be leveraged to further streamline the sufficient condition for vague convergence in our equivalence theorem.

EQUIVALENCE THEOREM III *Suppose $\{F_n\}$ and F are distributions with corresponding operators $\{\mathfrak{F}_n\}$ and \mathfrak{F} , respectively. Then $F_n \xrightarrow{v} F$ if, and only if, $\|\mathfrak{F}_n v - \mathfrak{F}v\| \rightarrow 0$ for every smooth function v .*

PROOF: Necessity follows via the equivalence theorem of the previous section as every smooth function u is necessarily a member of the class $\bar{\mathcal{C}}_b$ of continuous functions with finite limits at infinity. It suffices hence to prove sufficiency. Suppose that $\|\mathfrak{F}_n v - \mathfrak{F}v\| \rightarrow 0$ for every smooth v . It will be enough to show then that $\|\mathfrak{F}_n u - \mathfrak{F}u\| \rightarrow 0$ for every $u \in \bar{\mathcal{C}}_b$. Fix any $\epsilon > 0$ and any $u \in \bar{\mathcal{C}}_b$. Then there exists a smooth function v such that $\|u - v\| < \epsilon$. As

$$\mathfrak{F}_n u - \mathfrak{F}u = (\mathfrak{F}_n u - \mathfrak{F}_n v) + (\mathfrak{F}_n v - \mathfrak{F}v) + (\mathfrak{F}v - \mathfrak{F}u),$$

the triangle inequality yields

$$\begin{aligned} \|\mathfrak{F}_n u - \mathfrak{F}u\| &\leq \|\mathfrak{F}_n(u - v)\| + \|\mathfrak{F}_n v - \mathfrak{F}v\| + \|\mathfrak{F}(v - u)\| \\ &\leq \|\mathfrak{F}_n\| \cdot \|u - v\| + \|\mathfrak{F}_n v - \mathfrak{F}v\| + \|\mathfrak{F}\| \cdot \|u - v\|. \end{aligned}$$

Now \mathfrak{F}_n and \mathfrak{F} are operators corresponding to distributions so that $\|\mathfrak{F}_n\| = \|\mathfrak{F}\| = 1$. As v is smooth, by assumption $\|\mathfrak{F}_n v - \mathfrak{F}v\| \rightarrow 0$ and, in particular, $\|\mathfrak{F}_n v - \mathfrak{F}v\| < \epsilon$ for all sufficiently large n . It follows that $\|\mathfrak{F}_n u - \mathfrak{F}u\| < 3\epsilon$, eventually, and as $\epsilon > 0$ is arbitrary we must have $\|\mathfrak{F}_n u - \mathfrak{F}u\| \rightarrow 0$. ▶

In consequence, to establish vague convergence it will suffice to consider convergence for the class of smooth functions. This refined version of the equivalence theorem turns out to be the most useful for our purposes.

THE SEMIGROUP OF CONVOLUTIONAL OPERATORS

A little further investigation will help unlock the full power and elegance of the operator notation. We focus now on the subclass of convolutional operators.

Suppose \mathfrak{F} and \mathfrak{G} are the operators associated with the distributions F and G . It is natural to define the operator product \mathfrak{FG} to be the operator obtained by first applying \mathfrak{G} and then applying \mathfrak{F} , in notation, $\mathfrak{FG}u := \mathfrak{F}(\mathfrak{Gu})$. Written out explicitly, this means

$$\mathfrak{FG}u(t) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} u(t-x-y) dG(y) \right\} dF(x),$$

and it is clear that \mathfrak{FG} is the operator corresponding to the distribution $F * G$.

The operator product is a binary operation which implements a map from any pair of convolutional operators into another convolutional operator, $(\mathfrak{F}, \mathfrak{G}) \mapsto \mathfrak{FG}$. In the terminology of algebra, the family of convolutional operators equipped with a product forms a *semigroup*.

The operator \mathfrak{H}_0 corresponding to the Heaviside distribution H_0 concentrated at the origin satisfies $\mathfrak{H}_0 u = u$; it follows that \mathfrak{H}_0 serves the rôle of the *multiplicative identity* for the semigroup of convolutional operators: $\mathfrak{FH}_0 = \mathfrak{H}_0 \mathfrak{F} = \mathfrak{F}$.⁴

Operator products satisfy our ordinary intuition for multiplication. Indeed, operator products inherit associativity from convolution (or by Fubini's theorem if you must), $(\mathfrak{F}_1 \mathfrak{F}_2) \mathfrak{F}_3 = \mathfrak{F}_1 (\mathfrak{F}_2 \mathfrak{F}_3)$, and we may discard parenthetical grouping and write the product simply as $\mathfrak{F}_1 \mathfrak{F}_2 \mathfrak{F}_3$. Likewise, as convolution is commutative so is the operator product and $\mathfrak{FG} = \mathfrak{GF}$. In view of the commutative and associative properties of operator product, it is natural to identify the operator associated with the n -fold convolution $F * F * \cdots * F = F^{*n}$ as the operator power $\mathfrak{FF} \cdots \mathfrak{F} = \mathfrak{F}^n$.

Additivity of expectation yields one further familiar result: operator products distribute over sums, $\mathfrak{F}(\mathfrak{G}_1 + \mathfrak{G}_2) = \mathfrak{FG}_1 + \mathfrak{FG}_2$, as the reader can readily verify by writing out the integrals.

The following result is not only fundamental but is remarkable in its algebraic character and beautifully illustrative of the simplicity that the operator approach can bring to bear.

⁴Semigroups with an identity element are called *monoids*. The semigroup of convolutional operators is a *multiplicative monoid*.

REDUCTIONIST THEOREM Suppose $\mathfrak{F}_1, \mathfrak{F}_2, \mathfrak{G}_1$, and \mathfrak{G}_2 are operators associated with probability distributions. Then

$$\|\mathfrak{F}_1\mathfrak{F}_2u - \mathfrak{G}_1\mathfrak{G}_2u\| \leq \|\mathfrak{F}_1u - \mathfrak{G}_1u\| + \|\mathfrak{F}_2u - \mathfrak{G}_2u\|$$

for each bounded, continuous u .

PROOF: As operator products are commutative and distribute over addition, we obtain

$$\mathfrak{F}_1\mathfrak{F}_2 - \mathfrak{G}_1\mathfrak{G}_2 = (\mathfrak{F}_1 - \mathfrak{G}_1)\mathfrak{F}_2 + \mathfrak{G}_1(\mathfrak{F}_2 - \mathfrak{G}_2) = \mathfrak{F}_2(\mathfrak{F}_1 - \mathfrak{G}_1) + \mathfrak{G}_1(\mathfrak{F}_2 - \mathfrak{G}_2).$$

The triangle inequality now shows that

$$\begin{aligned} \|\mathfrak{F}_1\mathfrak{F}_2u - \mathfrak{G}_1\mathfrak{G}_2u\| &\leq \|\mathfrak{F}_2(\mathfrak{F}_1 - \mathfrak{G}_1)u\| + \|\mathfrak{G}_1(\mathfrak{F}_2 - \mathfrak{G}_2)u\| \\ &\leq \|\mathfrak{F}_2\| \cdot \|(\mathfrak{F}_1 - \mathfrak{G}_1)u\| + \|\mathfrak{G}_1\| \cdot \|(\mathfrak{F}_2 - \mathfrak{G}_2)u\|. \end{aligned}$$

The claimed result follows as \mathfrak{F}_2 and \mathfrak{G}_1 are operators associated with distributions, hence have norm 1. ▶

Induction quickly allows us to establish the general result: if $\mathfrak{F}_1, \dots, \mathfrak{F}_n$ and $\mathfrak{G}_1, \dots, \mathfrak{G}_n$ are convolutional operators then

$$\|\mathfrak{F}_1 \cdots \mathfrak{F}_n u - \mathfrak{G}_1 \cdots \mathfrak{G}_n u\| \leq \sum_{k=1}^n \|\mathfrak{F}_k u - \mathfrak{G}_k u\|$$

and a fortiori, if \mathfrak{F} and \mathfrak{G} are convolutional operators then

$$\|\mathfrak{F}^n u - \mathfrak{G}^n u\| \leq n \|\mathfrak{F}u - \mathfrak{G}u\|.$$

In spite of its deceptive algebraic simplicity, the result is worth more than a passing glance. On the left, an expression like $\|\mathfrak{F}^n u - \mathfrak{G}^n u\|$ conceals a formidable integration over n dimensions as the distributions F^{*n} and G^{*n} are induced from the n -fold product measures $F \otimes \cdots \otimes F = F^{\otimes n}$ and $G \otimes \cdots \otimes G = G^{\otimes n}$, respectively. On the right, however, the expression has simplified, remarkably, into an integration in one dimension involving only the univariate distributions F and G . While it is possible that one could discover such an elegant result by working directly with expectation with respect to an n -variate distribution, it is certainly unlikely—the structure of integration obscures the essential point of the underlying normed space, a point brought into vivid relief by the operator notation. As we see in the following chapter, this systematic reduction of an n -variate integral to a consideration of a one-dimensional integral is key to the proof of the central limit theorem.

6 A selection theorem

As we saw in Example 1.6, a sequence of distribution functions may converge pointwise to a limiting function, but the limit itself may not be a bona fide distribution function. The simplest example of this type is furnished by the sequence $F_n(x) = F(x + n)$ where F is any distribution function. In this case F_n converges pointwise to 0. It is useful to expand our definition of vague convergence to take in situations like this.

The problem with the above example is that the limit has a “mass at infinity”. We recall that a random variable X , or its distribution F , is *defective* if it has positive probability of taking the values $\pm\infty$, that is to say, if $F(-\infty, \infty) = P\{-\infty < X < \infty\} < 1$. Occasionally we will be pedantic and refer to the ordinary kind of distribution as *proper* to draw a clear distinction with their defective brethren though, strictly speaking, this is not necessary; a distribution function without qualification is assumed to be proper. We now allow our definition of vague convergence to include possibly defective limits: a sequence of distributions $\{F_n\}$ converges vaguely to a possibly defective distribution F if $F_n(\mathbb{I}) \rightarrow F(\mathbb{I})$ for every bounded interval of continuity \mathbb{I} of F . (We will reserve the terminology of convergence in distribution, however, only for situations where the limiting random variable is not defective.) If the limiting distribution F is defective we say that the convergence is *improper*; otherwise we say that the convergence is *proper*. We will only need to keep the distinction between proper and improper convergence for the rest of this section; limiting distributions will be resolutely proper elsewhere. As discussed at the end of Section 1, vague convergence will be proper if, and only if, for every $\epsilon > 0$, there exists a such that $F_n(-a, a) > 1 - \epsilon$ for all sufficiently large n .

We will need one concept from analysis. Suppose \mathbb{I} is any interval and \mathbb{D} any set of real numbers. In vivid terminology, we say that \mathbb{D} is *dense* in \mathbb{I} if, for every $\epsilon > 0$ and every $x \in \mathbb{I}$, the interval $(x - \epsilon, x + \epsilon)$ contains a point of \mathbb{D} other than x . If \mathbb{D} is dense in the real line then it is said to be *everywhere dense*. Thus, for instance, the set of rational numbers \mathbb{Q} is everywhere dense.

Many fundamental results depend upon our ability to pick a convergent subsequence from an *arbitrary* sequence of distributions. The archetypal result in this regard is due to E. Helly.⁵

HELLY'S SELECTION PRINCIPLE *Every sequence of probability distributions contains a subsequence which converges vaguely to a (possibly defective) limit distribution.*

PROOF: The key element of the proof is the elementary result from analysis known as the Bolzano–Weierstrass theorem: every bounded sequence has a convergent subsequence (the reader will find a proof in Section XXI.1 in the

⁵E. Helly, “Über lineare Funktionaloperationen”, *SitzBer. Akad. Wiss. Wien*, vol. 121, pp. 265–297, 1912.

Appendix). Suppose $\{F_n, n \geq 1\}$ is a sequence of probability distributions. Let $\{x_j, j \geq 1\}$ be any sequence of everywhere dense numbers. Then $F_1(x_j), F_2(x_j), \dots$ forms a sequence of numbers for each j . We claim that *there exists a subsequence of $\{F_n\}$ which converges at each point x_j* . The idea is to recursively build up an increasingly refined subsequence family via the flexible diagonalisation procedure introduced by G. Cantor.

As a *base* for a recurrence, begin with the sequence of numbers $F_1(x_1), F_2(x_1), \dots$. Clearly, these numbers are bounded between 0 and 1 as the F_n are probability distributions. It follows by the Bolzano–Weierstrass theorem that there exists a subsequence $\{F_{11}, F_{12}, \dots\}$ of the original sequence of distributions $\{F_1, F_2, \dots\}$ such that $\{F_{1n}(x_1), n \geq 1\}$ converges.

To establish the *recurrence*, for $j \geq 1$, given a subsequence of distributions $\{F_{j1}, F_{j2}, \dots\}$ such that the sequence of numbers $\{F_{jn}(x_j)\}$ converges, select a further subsequence $\{F_{j+1,1}, F_{j+1,2}, \dots\}$ such that $\{F_{j+1,n}(x_{j+1})\}$ converges. The existence of such a subsequence of $\{F_{jn}\}$ is again guaranteed by the Bolzano–Weierstrass theorem.

It is now an easy matter of induction to see that $\{F_{jn}\}$ is a subsequence of each of the preceding subsequences of distributions $\{F_{in}\}$ for $i < j$. As any subsequence of a convergent sequence converges, and to the same limit, it follows that, for each j , the subsequence of distributions $\{F_{jn}\}$ converges at each of the points x_1, x_2, \dots, x_j . Cantor's diagonal method now allows us to identify a subsequence that converges simultaneously at *all* of the points x_j . Set $G_n = F_{nn}$ for each n . Then $\{G_1, G_2, \dots\}$ is a subsequence of $\{F_{11}, F_{12}, \dots\}$ and hence converges at the point x_1 . Likewise, for any j , $\{G_j, G_{j+1}, \dots\}$ is a subsequence of $\{F_{j1}, F_{j2}, \dots\}$ and hence converges at the point x_j . (The preamble of the first $j - 1$ values $G_1(x_j), \dots, G_{j-1}(x_j)$ is immaterial.) It follows that $\{G_1, G_2, \dots\}$ is a subsequence of $\{F_1, F_2, \dots\}$ that converges at each of the points x_1, x_2, \dots . The idea is illustrated in Figure 2.

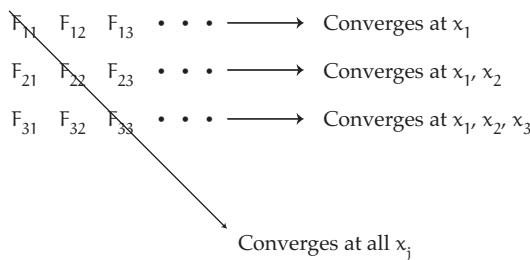


Figure 2: A selection principle.

For each j , write $G(x_j)$ for the limit of the sequence $G_n(x_j)$. For x not in the set $\{x_j\}$, we may complete the definition of the function G by setting $G(x)$ to be the greatest lower bound of the $G(x_j)$ where x_j exceeds x , in notation,

$G(x) = \inf\{G(x_j) : x_j > x\}$. The resulting function is bounded between 0 and 1 and increasing but is not necessarily a distribution function (defective or proper) as it may not be right continuous. It is a simple matter to fix it so that it is: set $F(x) = G(x)$ at all points of continuity of G , and at points of jump of G , set $F(x) = G(x+)$. Then F is a (possibly defective) distribution function. All that remains to be done is to verify that F is indeed the limit of the subsequence $\{G_n\}$ at all points of continuity of F .

Suppose x is any point of continuity of F . Then $F(x) = G(x)$. If x is in the set $\{x_j\}$ then there is nothing to prove: $G_n(x) \rightarrow G(x)$ by construction. Suppose now that x is not in $\{x_j\}$. Fix any $\epsilon > 0$. As x is a point of continuity of G we can find points a and b in the set $\{x_j\}$ with $a < x < b$ and so that $0 \leq G(b) - G(a) < \epsilon$. As G_n is a distribution function it is monotone for each value of n , hence $G_n(a) \leq G_n(x) \leq G_n(b)$. Now $G_n(a) \rightarrow G(a)$ and $G_n(b) \rightarrow G(b)$ whence, for every $\epsilon > 0$, $0 \leq G_n(b) - G_n(a) < 3\epsilon$, eventually, via the triangle inequality. It follows that $G_n(x) - 4\epsilon < G(x) < G_n(x) + 4\epsilon$ for all sufficiently large n , and thus, $G_n(x) \rightarrow G(x)$ at all points of continuity of G .

We've hence exhibited a subsequence of distributions $\{G_n\}$ and a possibly defective distribution F such that $\{G_n\}$ converges to F at all points of continuity, or, what is the same thing, $\{G_n\}$ converges vaguely to F . ►

An immediate consequence of the selection principle is the important theorem of Arzelà and Ascoli in analysis.⁶

THE ARZELÀ–ASCOLI THEOREM *Suppose $\{u_n\}$ is a bounded, equicontinuous sequence of functions. Then there exists a subsequence $\{u_{n_i}\}$ which converges to a continuous limit function u , the convergence being uniform in every closed and bounded interval.*

PROOF: Suppose that the sequence $\{x_j\}$ is everywhere dense. It is clear that the diagonalisation argument in the proof of Helly's selection principle is not restricted to distribution functions but works for any bounded sequence. It follows that there exists a subsequence $\{u_{n_1}, u_{n_2}, \dots\}$ which converges at each point x_j . The equicontinuity property of the sequence suggests that the u_{n_i} will approach limiting values uniformly in a neighbourhood of any of the x_j . Indeed, suppose $x \notin \{x_j\}$. By the triangle inequality,

$$|u_{n_r}(x) - u_{n_s}(x)| \leq |u_{n_r}(x) - u_{n_r}(x_j)| + |u_{n_r}(x_j) - u_{n_s}(x_j)| + |u_{n_s}(x_j) - u_{n_s}(x)| \quad (6.1)$$

for each x_j . Recall that equicontinuous means that for every n and every $\epsilon > 0$ there exists $\delta = \delta(\epsilon) > 0$ independent of n such that $|u_n(x) - u_n(y)| < \epsilon$ whenever $|x - y| < \delta$. As $\{u_{n_i}\}$ is a subsequence of $\{u_n\}$, by equicontinuity, there exists x_j such that

$$|u_{n_i}(x) - u_{n_i}(x_j)| < \epsilon \quad (6.2)$$

for all i . With x_j chosen appropriately then, the first and third terms on the right of (6.1) are less than ϵ . As $\{u_{n_i}(x_j)\}$ converges, there exists m such that $|u_{n_r}(x_j) - u_{n_s}(x_j)| < \epsilon$

⁶C. Arzelà, "Sulle funzioni di linee", *Mem. Accad. Sci. Ist. Bologna Cl. Sci. Fis. Mat.*, vol. 5, no. 5, pp. 55–74, 1895; G. Ascoli, "Le curve limiti di una varietà data di curve", *Atti della R. Accad. Dei Lincei Memorie della Cl. Sci. Fis. Mat. Nat.*, vol. 18, no. 3, pp. 521–586, 1883–1884.

for all values of r and s greater than m . Thus, $|u_{n_r}(x) - u_{n_s}(x)| < 3\epsilon$, eventually. In other words, $\{u_{n_i}(x)\}$ is a Cauchy sequence, hence convergent. Thus $\{u_{n_i}\}$ converges everywhere and we may set $u = \lim u_{n_i}$. It follows from (6.2) that u is continuous.

If we restrict attention to values x in a closed and bounded interval \mathbb{I} then it will suffice to consider a finite number of x_j s for which (6.2) holds. Indeed, select $\delta = \delta(\epsilon)$ independent of x and n by the equicontinuity condition. As any bounded interval can be covered by a finite number of subintervals of length δ , and as $\{x_j\}$ is everywhere dense, we may select one x_j per subinterval. With the x_j s restricted to a finite set the right-hand side of (6.1) will be bounded above by 3ϵ uniformly over the interval \mathbb{I} for all r and s sufficiently large. It follows that $\{u_{n_i}\}$ converges to u uniformly in \mathbb{I} . ▶

While I have presented the result in one dimension for clarity, the proof carries over to any number of dimensions; simply strike out scattered references to intervals and replace them by rectangles and interpret distance between points in the usual Euclidean sense.

A family of continuous functions defined on a closed and bounded set is said to be *compact* if every sequence of these functions contains a uniformly convergent subsequence. In view of the above, we may identify compactness in the space of continuous functions with closed, bounded, and equicontinuous. It's not very intuitive yet at this level, but it grows upon one—the Arzelà–Ascoli theorem is fundamental in the analysis of the space of continuous functions.

The following theorem shows us how to extend the equivalence theorem of Section 2 to accommodate our expanded definition of vague convergence. An additional useful feature is that it is only necessary to know that the limit exists, not what it is.

EQUIVALENCE THEOREM IV *A sequence of probability distributions $\{F_n\}$ converges vaguely to a (possibly defective) limit if, and only if, the sequence of values $\int u dF_n$ converges for every continuous function u vanishing at infinity.*

PROOF: To prove necessity, suppose $F_n \overset{v}{\rightarrow} F$ for some F , possibly defective. Let u be any continuous function vanishing at infinity and pick any $\epsilon > 0$. Then we can pick an interval $\mathbb{I} = (a, b)$ such that $|u(x)| < \epsilon$ for all $x \in \mathbb{I}^c$. Leveraging the triangle inequality, we then obtain

$$\left| \int u dF_n - \int u dF \right| \leq \left| \int_{\mathbb{I}} u dF_n - \int_{\mathbb{I}} u dF \right| + 2\epsilon.$$

The rest of the proof follows that for proper limit distributions verbatim.

The tricky bit in proving sufficiency is that the limiting distribution is not specified.⁷ No matter, this is where Helly's selection principle kicks in: there

⁷The reader conversant with functional analysis will note that sufficiency follows directly because $\lim \int u dF_n$ is a linear functional whence it can be written as the expectation of u with respect to some F by the Riesz representation theorem.

exists a subsequence $\{F_{n_i}, i \geq 1\}$ which converges vaguely to a limit, say, F , possibly defective. Suppose now that $\int u dF_n$ is convergent for any given continuous u vanishing at infinity. As any subsequence of a convergent sequence also converges, and to the same limit, it follows that $\int u dF_{n_i}$ is convergent with the same limit point. But $F_{n_i} \xrightarrow{v} F$ whence, by the just-proved necessity condition, $\int u dF_{n_i} \rightarrow \int u dF$, and perforce, $\int u dF$ must be the limit point of the sequence $\int u dF_n$ as well. We now merely have to follow along with the sandwich argument in the proof of the equivalence theorem for proper limit distributions to finish up. ►

The selection principle espoused in Helly's theorem has a scope beyond the realm of probability and has a wide range of applications. I will provide a mere soupçon to whet the appetite.



7 Two by Bernstein

The reader is well aware that a function has a convergent Taylor series if it is smooth enough. S. N. Bernstein provided the following sufficient condition.⁸

THEOREM 1 *Suppose f is positive and has positive derivatives of all orders on $0 \leq x < 1$. Then f has a convergent power series of the form $f(x) = p(0) + p(1)x + p(2)x^2 + \dots + p(k)x^k + \dots$ where the coefficients $p(k) = f^{(k)}(0)/k!$ are positive for each k .*

The usual proof deals directly with the remainder term in a Taylor expansion but Helly's selection principle provides an elegant alternative.

We begin with the *difference operator* Δ encountered previously in Section XVI.5. With $h > 0$ a fixed constant, the operator is defined by $\Delta f(x) = f(x+h) - f(x)$. As before, its iterates Δ^k are defined recursively by $\Delta^k f(x) = \Delta(\Delta^{k-1} f(x))$ for $k \geq 1$, where, by convention, we set $\Delta^0 f(x) = f(x)$.

LEMMA *If f is positive and has positive derivatives of all orders then $\Delta^k f \geq 0$ for each $k \geq 0$.*

PROOF: Suppose as induction hypothesis that, for some k , any positive function g with positive derivatives of all orders has positive difference iterates $\Delta^0 g, \Delta^1 g, \dots, \Delta^k g \geq 0$. Suppose $f \geq 0$ has positive derivatives of all orders. Writing $g_j = \Delta^j f$, we have

$$g_{k+1}(x) = \Delta g_k(x) = g_k(x+h) - g_k(x) = \int_x^{x+h} g'_k(t) dt.$$

Now, $g'_1(t) = \frac{d}{dt} \Delta f(t) = f'(t+h) - f'(t) = \Delta f'(t)$ by linearity of differentiation, and, working inductively, we obtain $g'_k(t) = \frac{d}{dt} \Delta^k f(t) = \Delta^k f'(t)$. It follows that $g_{k+1}(x) = \int_x^{x+h} \Delta^k f'(t) dt$. As f' is positive and has positive derivatives of all orders, by induction hypothesis $\Delta^k f'(t) \geq 0$ for all t . Thus, $\Delta^{k+1} f(x) = g_{k+1}(x) \geq 0$, completing the induction. ►

⁸S. N. Bernstein, "Sur la définition et les propriétés des fonctions analytiques d'une variable réelle", *Mathematische Annalen*, vol. 75, pp. 449–468, 1914.

We recall that, in the language and terminology introduced in Section XV.1, if F is a distribution concentrated on the positive half-line then its Laplace transform is defined by $\widehat{F}(\zeta) = \int_{\mathbb{R}^+} e^{-\zeta x} dF(x)$ for $\zeta \geq 0$, and, if F is arithmetic and assigns mass $p(k)$ to the positive integer k , then, by setting $x = e^{-\zeta}$, we obtain the generating function $\widehat{F}(-\log x) = \sum_{k \geq 0} p(k)x^k$ for $0 < x \leq 1$ characteristic of the distribution F .

PROOF OF THEOREM 1: With preliminaries concluded, suppose now that, for each $k \geq 0$, $f^{(k)}(x) \geq 0$ in the interval $0 \leq x < 1$, with the usual convention $f^{(0)} = f$ in force. Then f is positive and increasing. Suppose, to begin, that f is bounded and, specifically, $f(1-) = 1$. We may then extend f to a continuous function on the closed unit interval $[0, 1]$. Restating (XVI.5.3) for convenience, we see that the Bernstein polynomials

$$f_n(x) = \sum_{j=0}^n f(jn^{-1}) \binom{n}{j} x^j (1-x)^{n-j} = \sum_{k=0}^n \binom{n}{k} \Delta^k f(0) x^k$$

have positive coefficients $p_n(k) = \binom{n}{k} \Delta^k f(0) \geq 0$ in view of our lemma. Allowing x to tend to one, we observe that $1 = f(1) = \sum_{k=0}^n p_n(k)$ whence, for each n , the values $\{p_n(k)\}$ determine an arithmetic probability distribution $F_n(t) = \sum_k p_n(k) H_0(t - k)$ concentrated on the positive integers. It follows that, for each n , the Bernstein polynomial $f_n(x) = \sum_k p_n(k)x^k = \widehat{F}_n(-\log x)$ represents the generating function of the probability distribution F_n . Specialised to these circumstances, Bernstein's proof of Weierstrass's theorem outlined in Section XVI.5 says that the sequence of generating functions $\{f_n, n \geq 0\}$ converges uniformly to the limit function f .

On the other hand, by Helly's selection principle, corresponding to the sequence of probability distributions $\{\mathbb{F}_n, n \geq 0\}$ there exists a convergent subsequence $\{\mathbb{F}_{n_j}, j \geq 1\}$ converging vaguely to a limit distribution $F(t) = \sum_k p(k) H_0(t - k)$, possibly defective, concentrated on the positive integers. Here, the values $p(k)$ are positive and $\sum_{k=0}^{\infty} p(k) \leq 1$. For any $0 \leq x < 1$ we may now identify $u(t) = x^{|t|}$ in the equivalence theorem IV of the previous section to conclude that

$$f_{n_j}(x) = \sum_{k \geq 0} p_{n_j}(k)x^k = \int_{\mathbb{R}^+} x^t d\mathbb{F}_{n_j}(t) \rightarrow \int_{\mathbb{R}^+} x^t dF(t) = \sum_{k \geq 0} p(k)x^k = f^*(x).$$

The subsequence of generating functions $\{f_{n_j}, j \geq 1\}$ hence converges to a limit function f^* which may be identified as the generating function of a (possibly defective) distribution F concentrated on the positive integers. But as $f_n \rightarrow f$ and any subsequence $\{f_{n_j}\}$ of the convergent sequence $\{f_n\}$ must have the same limit, it follows that $f(x) = f^*(x) = \sum_{k \geq 0} p(k)x^k$ may be expressed as a power series with positive coefficients. Identifying $f(x) = \widehat{F}(\log x)$ with the generating function of the distribution F , the proof of Theorem XV.1.3 may be adapted almost verbatim to conclude that, for each $k \geq 0$, we may differentiate termwise under the summation to obtain

$$f^{(k)}(x) = p(k)k! + p(k+1)(k+1)!x + p(k+2)(k+2)!x^2 + \dots,$$

the series convergent for all $0 \leq x < 1$. By setting $x = 0$, we obtain $f^{(k)}(0) = k!p(k)$ and we recover the familiar coefficients of the Taylor series. In particular, we conclude that the power series expansion of f is unique.

We may relax the requirement that $f(1-) = 1$ for a bounded f by working instead with the function $\tilde{f}(x) = f(x)/f(1-)$. The preceding argument works without change for the normalised function \tilde{f} whence f has a convergent power series whose coefficients are obtained by multiplying those of \tilde{f} by $f(1-)$.

The function $(1-x)^{-1}$ is a typical instance of an increasing function on $0 \leq x < 1$ with an unbounded limit as $x \uparrow 1$. But in cases such as this, we may work instead on the restriction f to the interval $[0, 1 - m^{-1}]$. If we set $\tilde{f}_m(x) = f((1 - m^{-1})x)/f(1 - m^{-1})$ then \tilde{f}_m is positive, increasing, and bounded on $[0, 1]$ with $\tilde{f}_m(1) = 1$. As $\tilde{f}_m^{(n)}(x) = (1 - m^{-1})^n f^{(n)}(x)/f(1 - m^{-1})$, it is clear that \tilde{f}_m has positive derivatives of all orders and hence has a convergent Taylor series expansion everywhere on $[0, 1]$. Consequently, $f(x) = f(1 - m^{-1})\tilde{f}_m(x/(1 - m^{-1}))$ has a convergent Taylor series expansion in the closed interval $[0, 1 - m^{-1}]$. By allowing m to tend to infinity, it follows that $f(x)$ has a convergent Taylor series for $0 \leq x < 1$, uniqueness carrying the day. ►

We recall per Section XV.3 that a positive function ψ is *completely monotone* if it has derivatives of all orders and they alternate in sign, that is, $(-1)^n \psi^{(n)}(\zeta) \geq 0$ for each $n \geq 0$. As usual, we identify $\psi^{(0)} = \psi$. In this language we may restate Theorem XV.2.3' compactly by saying that Laplace transforms of distributions are completely monotone. Bernstein proved a remarkable result in an unexpected direction.⁹

THEOREM 2 *A function ψ is completely monotone if, and only if, it is the Laplace transform of a distribution F .*

PROOF: Sufficiency was handled in Sections XV.1 and XV.2. To prove necessity, suppose ψ is completely monotone on the interval $(0, \infty)$ and suppose $\psi(0) = 1$. (If ψ is defined and bounded on the interval (ζ_0, ∞) and $\psi(\zeta_0)$ is not necessarily unit, we may consider the scaled translate $\psi(\zeta - \zeta_0)/\psi(\zeta_0)$ to reduce it to the considered case; the unbounded case may be handled by normalisation as in the proof of the general inversion theorem for the Laplace transform of the distributions of measures in Section XV.2.) Then the function $f_m(s) = \psi(m(1 - s))$ has positive derivatives of all orders for $0 \leq s < 1$ and, by Theorem 1, has a convergent power series $f_m(s) = p_m(0) + p_m(1)s + p_m(2)s^2 + \dots$ where the coefficients $p_m(k)$ are positive and add to one. Setting $s = e^{-\zeta/m}$, we obtain

$$\psi_m(\zeta) := f_m(e^{-\zeta/m}) = \psi(m(1 - e^{-\zeta/m})) = \sum_{k \geq 0} p_m(k)e^{-k\zeta/m} \quad (\zeta > 0).$$

The values $\{p_m(k), k \geq 0\}$ determine a distribution $F_m(t) = \sum_{k \geq 0} p_m(k)H_0(t - k/m)$ concentrated on the points k/m for $k \geq 0$. It is clear hence that, for each m , $\psi_m(\zeta) = \widehat{F_m}(\zeta)$ represents the Laplace transform of the distribution $F_m(t)$. By Helly's selection principle, there exists a subsequence $\{F_{m_j}, j \geq 1\}$ converging vaguely to a (possibly defective) distribution F concentrated on the positive half-line and so, setting $u(x) = e^{-\zeta|x|}$ in the equivalence theorem, we have $\widehat{F_{m_j}}(\zeta) \rightarrow \widehat{F}(\zeta)$ for each $\zeta > 0$.

⁹S. N. Bernstein, "Sur les fonctions absolument monotones", *Acta Mathematica*, vol. 52, pp. 1–66, 1928.

On the other hand, for each fixed ζ , $m(1 - e^{-\zeta/m}) \rightarrow \zeta$ as $m \rightarrow \infty$. By the continuity of ψ it follows that $\psi_m(\zeta) \rightarrow \psi(\zeta)$ as $m \rightarrow \infty$. As any subsequence $\{\psi_{m_j}, j \geq 1\}$ must have the same limit, it must hence be the case that $\psi(\zeta) = \widehat{F}(\zeta)$ is the Laplace transform of a distribution F . The uniqueness theorem of Section XV.2 ensures that F is completely characterised by its Laplace transform. ▶

In the next section I will provide a much more concrete illustration of the use of the selection principle in a historical context. Our walk down memory lane culminates with an elegant probabilistic resolution of a famous theorem in number theory.



8 Equidistributed numbers, from Kronecker to Weyl

Given a real number ϑ , what can be said about the set of its integer multiples $\{n\vartheta : \text{integer } n\}$? More concretely, for any given x , does there exist an integer n for which $x - n\vartheta$ is nearly an integer?

The problem, as stated, is not very interesting if ϑ is rational, say equal to a/b in lowest terms, for then $n\vartheta$ will differ only by an integer amount from one of the b values $0, 1/b, \dots, (b-1)/b$. We are therefore led to consider *irrational* ϑ .

By reducing numbers modulo one, we can recast the problem without irritating integer shifts. Solely for the purposes of this section and the next, for each real x introduce the nonce notation $(x) := x \bmod 1$ for the *fractional part* of x . Alternatively, $(x) = x - [x]$ with $[x]$ denoting the integer part of x in standard notation. Clearly, $0 < (x) < 1$ whenever x is not an integer [the case of integer x is trivial and in this case $(x) = 0$]. In these terms, the question is whether, for any given $x \in (0, 1)$, there is an integer n such that $x - (n\vartheta)$ is arbitrarily small. The answer, and more besides, was provided by Kronecker in a *tour de force* in 1884.¹⁰ The theorem articulated below is a specialisation of his result to one dimension, in which form it may have been known to Chebyshev. The pretty geometrical proof is due to G. H. Hardy and J. E. Littlewood.

KRONECKER'S THEOREM *If ϑ is irrational then the sequence $\{(n\vartheta) : \text{integer } n\}$ is dense in the unit interval $(0, 1)$.*

PROOF: We consider points in the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, i.e., the real line modulo one, obtained by winding up the real line and identifying the points $x, x \pm 1, x \pm 2, \dots$. We can then think of any point x on the line as the *name* of a point on the circle \mathbb{T} . This is the same artifice we encountered in Section IX.2.

We may identify $\mathbb{D} := \{(n\vartheta), n \in \mathbb{Z}\}$ with a set of points in \mathbb{T} . Let $\widehat{\mathbb{D}}$ denote the set of *accumulation points* of \mathbb{D} on the circle, that is to say, the set of points t in \mathbb{T} for which there are points of \mathbb{D} other than t in every open interval containing t . The points of \mathbb{D} are either isolated or themselves accumulation points. The assertion of the theorem is equivalent to saying that every point in the circle \mathbb{T} is an accumulation point of \mathbb{D} .

Write $\text{cl } \mathbb{D} = \mathbb{D} \cup \widehat{\mathbb{D}}$ for the closure of \mathbb{D} . We begin by showing that $\text{cl } \mathbb{D}$ includes all the points in the circle \mathbb{T} . Suppose x does not belong to $\text{cl } \mathbb{D}$. Then there is an interval I around x , say, $(x - a, x + b)$ with positive a and b , which contains no point of \mathbb{D} . Say

¹⁰L. Kronecker, *Berliner Sitzungs-berichte* [Werke, iii(i), pp. 47–110], 1884.

that this is an *excluded* interval. As the circumference of the circle is bounded, among all such excluded intervals there must be a *largest*. A formal argument may be sketched as follows. For any interval \mathbb{I} , let $\lambda(\mathbb{I})$ denote its measure (or length). Write $\ell := \sup \lambda(\mathbb{I})$ where the supremum is over all excluded intervals \mathbb{I} . As, by assumption, there is at least one excluded interval, we must have $\ell > 0$. Then there is a sequence of excluded intervals \mathbb{I}_n of positive length such that $\lambda(\mathbb{I}_n) \rightarrow \ell$. In particular, there exists N such that for all $n \geq N$, $\ell/2 \leq \lambda(\mathbb{I}_n) \leq \ell$. To each such excluded interval \mathbb{I}_n there exists a largest excluded interval \mathbb{I}'_n containing \mathbb{I}_n ; this is simply the union of all the excluded intervals containing \mathbb{I}_n . Then we have $\ell/2 \leq \lambda(\mathbb{I}_n) \leq \lambda(\mathbb{I}'_n) \leq \ell$ for all $n \geq N$ and, in consequence, $\lambda(\mathbb{I}'_n) \rightarrow \ell$ from below. Now, any two of the maximal excluded intervals \mathbb{I}'_m and \mathbb{I}'_n must either coincide or be disjoint. For if they overlap then a larger excluded interval is created and this is precluded by the maximality of each of \mathbb{I}'_m and \mathbb{I}'_n . As there can be at most $\lceil 2/\ell \rceil$ disjoint intervals each of length $\geq \ell/2$ in \mathbb{T} , it follows that for $n \geq N$, the maximal excluded intervals \mathbb{I}'_n range over only a finite range of distinct excluded intervals. Suppose none of these has length equal to ℓ . Then the largest of these maximal excluded interval lengths is some quantity $\ell' < \ell$. But then $\max_{n \geq N} |\lambda(\mathbb{I}'_n) - \ell| \geq \ell - \ell' > 0$ and $\lambda(\mathbb{I}'_n)$ stays bounded away from ℓ . Contradiction.

Accordingly, suppose x_0 is the midpoint of a maximal length excluded interval $\mathbb{I}_{(x_0)} := (x_0 - \ell/2, x_0 + \ell/2)$. It follows that $|x_0 - n\vartheta| \geq \ell/2$ for each n . But then the point $x_0 - \vartheta$ is also surrounded by a maximal length excluded interval $\mathbb{I}_{(x_0-\vartheta)} := (x_0 - \vartheta - \ell/2, x_0 - \vartheta + \ell/2)$ and, arguing in this fashion, we create a sequence of excluded intervals $\mathbb{I}_{(x_0)}, \mathbb{I}_{(x_0-\vartheta)}, \mathbb{I}_{(x_0-2\vartheta)}, \dots$, each of which has positive length ℓ . These intervals cannot coincide as ϑ is irrational and cannot overlap for if any two did, their union would create a larger excluded interval which is precluded by assumption. But it is impossible for such a sequence of intervals to exist as the circumference of the circle is unit and it cannot accommodate an infinity of disjoint intervals of equal positive length. It follows that there can be no excluded interval \mathbb{I} and this implies that each x must lie in $\text{cl } \mathbb{D}$.

It only remains to check that \mathbb{D} does not have any isolated points $x \in \mathbb{D} \setminus \widehat{\mathbb{D}}$. But if such a point exists then there is an interval around x which contains no points of \mathbb{D} , except x itself, and therefore there are points in the neighbourhood of x that belong neither to \mathbb{D} nor to $\widehat{\mathbb{D}}$. But as all points in the circle \mathbb{T} lie in $\text{cl } \mathbb{D} = \mathbb{D} \cup \widehat{\mathbb{D}}$, this possibility is obviated. It follows that each x must, in fact, be an accumulation point of \mathbb{D} . ▶

Kronecker's theorem, while deep, does not tell the entire story. What is left open is the issue of how the numbers $(n\vartheta)$ are distributed around the circumference of the unit circle. The answer was provided by H. Weyl in a landmark paper in 1916.¹¹

WEYL'S EQUIDISTRIBUTION THEOREM *Suppose ϑ is irrational. Let \mathbb{I} be any subinterval (or more generally, Borel set) of the unit interval and let $\lambda(\mathbb{I})$ be its measure (or length). Then $\frac{1}{n} \text{card}\{(k\vartheta) \in \mathbb{I} : 1 \leq k \leq n\} \rightarrow \lambda(\mathbb{I})$. In particular, the asymptotic fraction of the numbers $(n\vartheta)$ that fall in any interval (a, b) with $0 < a < b < 1$ tends to the interval length $b - a$.*

PROOF: For each n , let F_n be the discrete distribution which places equal mass $1/n$ at each of the points $(\vartheta), (2\vartheta), \dots, (n\vartheta)$ on the circle \mathbb{T} . By Helly's selection principle, there

¹¹H. Weyl, "Über die Gleichverteilung von Zahlen mod. Eins.", *Mathematische Annalen*, vol. 77, pp. 313–352, 1916.

exists a subsequence $\{F_{n_i}, i \geq 1\}$ which converges vaguely to a limiting distribution F on \mathbb{T} . This F cannot be defective as \mathbb{T} is bounded and hence the convergence is proper. Write $\{\mathfrak{F}_{n_i}, i \geq 1\}$ and \mathfrak{F} for the associated convolutional operators.

We now adapt the equivalence theorem of Section 3 to the space of continuous functions defined on the circle \mathbb{T} . (These functions are automatically bounded as they are continuous over a closed and bounded, i.e., *compact*, set.) As $F_{n_i} \xrightarrow{v} F$, we must have $\mathfrak{F}_{n_i} u \rightarrow \mathfrak{F}u$ for every continuous function on \mathbb{T} . In detail,

$$\mathfrak{F}_{n_i} u(t) = \frac{1}{n_i} \sum_{k=1}^{n_i} u(t - k\vartheta) \rightarrow \int_{\mathbb{T}} u(t - x) dF(x) = \mathfrak{F}u(t).$$

The left-hand side is unaffected by the replacement of t by $t - \vartheta$. It follows by induction that the right-hand side satisfies $\mathfrak{F}u(t) = \mathfrak{F}u(t - k\vartheta)$ for all integers k . But by Kronecker's theorem, the sequence $\{k\vartheta, k \geq 1\}$ is dense in \mathbb{T} . As $\mathfrak{F}u(t)$ is continuous it follows that it must be equal to a constant everywhere on the circle. We have thus shown that the convolution $\mathfrak{F}u$ is a constant for every continuous u on the circle. This is only possible if the limiting distribution F is the *uniform distribution on the circle* that we had encountered in Section IX.2: F assigns to any interval \mathbb{I} its length, $F(\mathbb{I}) = \lambda(\mathbb{I})$. Indeed, suppose u is the indicator for an interval \mathbb{I} of a given length ℓ . While we cannot directly apply our conclusion to this u because it is discontinuous, the Lévy sandwich in the proof of the equivalence theorem of Section 2 shows that we can apply the conclusion to the continuous functions u_{\pm} which sandwich u . It follows that $\mathfrak{F}u$ is bounded between constant values $\mathfrak{F}u_-$ and $\mathfrak{F}u_+$ and the latter two values can be made arbitrarily close. It follows that $\mathfrak{F}u$ is a constant which is the same as saying that $F(\mathbb{I}) = F(\mathbb{I} + t)$ for all affine shifts t on the circle. We conclude therefore that $F(\mathbb{I})$ is a constant determined only by the length of \mathbb{I} and not its position. In particular, it follows by additivity that $F[0, 1] = 1$, $F[0, 1/2] = F[1/2, 1] = 1/2$, $F[0, 1/3] = F[1/3, 2/3] = F[2/3, 1] = 1/3$, and in general, that $F[0, 1/n] = 1/n$. Thus, $F(\mathbb{I}) = 1/n$ for any interval of length $1/n$. By additivity, if \mathbb{I} is any interval of rational length, say, m/n , then $F(\mathbb{I}) = m/n$. We extend this argument to intervals with irrational length by continuity of probability measure whence we finally obtain that, for any interval \mathbb{I} of length ℓ , we must have $F(\mathbb{I}) = \ell$. By uniqueness, F must perforce be the uniform distribution on the circle. As it is clear that no other limit is possible, it follows that F is the vague limit of the entire sequence $\{F_n\}$. ▶

If the reader now revisits Chapter VI, and in particular Problem VI.8, she will see the value of Lévy's sandwich argument from a different vantage point and an alternative, essentially Fourier-theoretic, proof of Weyl's theorem.

Weyl's theorem says that the numbers $(n\vartheta)$ are distributed in fair and just fashion uniformly across the unit interval. In conjunction with Borel's law of normal numbers in Section V.7, the reader may take Weyl's equidistribution theorem as providing additional evidence that numbers play fair.

Weyl's theorem had an enormous impact. T. W. Körner has pointed out, for instance, that the *reviews* alone of papers springing from Weyl's theorem fill over 100 pages in Chapter K of *Mathematics Reviews in Number Theory 1940–1972*. What is noteworthy in our exposition is how elegantly the selection theorem resolves the problem when it is put in a probabilistic perspective. To gain an appreciation of how difficult the problem is viewed from a purely number-theoretic viewpoint abeyant the probabilistic

angle, the reader may wish to consult Hardy and Wright who devote an entire chapter to the problem in their classical work on the theory of numbers.¹² In conjunction with Chapters V, VI and Sections XVI.6, 7, our analysis reinforces the slogan that understanding and exploring the link between number theory and probability richly enhances and contributes to both disciplines.

9 Walking around the circle

Suppose X is a random variable with distribution F , and X_1, X_2, \dots are independent copies of X . Consider the sums $S_n = X_1 + \dots + X_n$. If F has finite variance the central limit theorem explored in the following chapter tells us that a suitably centred and scaled version of S_n converges in distribution to the normal. What, if anything, can we say about the *fractional part* of S_n ?

As in the previous section, write $(S_n) = S_n \bmod 1$ for the fractional part of S_n . As before, we may consider (S_n) to take values in the real line modulo one, i.e., the circle $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ obtained by winding up the real line and identifying the points $x, x \pm 1, x \pm 2, \dots$. Again, we may view any point x on the line as the name of a point on the circle. A collection of points viewed in the circle may have certain geometric attributes. We will without additional comment refer to such attributes at need with the understanding that in such cases the points are being viewed in their rôle in the circle.

Only the fractional parts of the X_k matter to (S_n) ; indeed, $(S_n) = ((X_1) + \dots + (X_n))$. It will be convenient therefore to consider F to be a distribution on the circle \mathbb{T} . Then (S_n) has distribution given by the n -fold convolution F^{*n} where all convolution integrals are interpreted as integrals over \mathbb{T} and all points are identified modulo one.

To get some understanding of the behaviour of F^{*n} let us consider some special cases. Suppose F is concentrated at the vertices of an equilateral triangle; without loss of generality we may assume that the points of concentration are 0, $1/3$, and $2/3$. It is easy to see now that each successive convolution with F (bear in mind that points are reduced modulo one) will return a distribution concentrated on the same three points. It follows that F^{*n} is also concentrated at 0, $1/3$, and $2/3$ for each value of n and nothing interesting transpires as the convolutions mount. This behaviour recurs whenever the set of concentration of F is contained in the vertices of a regular polygon. Remarkably, polygons are the only exceptional cases not leading to bland uniformity.

THEOREM *If F is not concentrated on the vertices of a regular polygon, then $\{F^{*n}, n \geq 1\}$ converges vaguely to the uniform distribution on the circle \mathbb{T} .*¹³

PROOF: Suppose u_0 is a continuous function on \mathbb{T} . Let u_n denote the iterates obtained by convolving u_0 repeatedly with respect to F , that is to say,

$$u_n(t) = \int_{\mathbb{T}} u_{n-1}(t-x) dF(x) = \int_{\mathbb{T}} u_0(t-x) dF^{*n}(x),$$

¹²G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers*, Chapter XXIII, pp. 375–293. Oxford: Oxford University Press, 1979.

¹³For generalisations to variable distributions see P. Lévy, *Bulletin de la Société Mathématique de France*, vol. 67, pp. 1–41, 1939.

or, in operator notation, $u_n = \mathfrak{F}u_{n-1} = \mathfrak{F}^n u_0$. As \mathbb{T} is closed and bounded, u_0 is uniformly continuous and in consequence the sequence of functions $\{u_n\}$ is equicontinuous. By the Arzelà–Ascoli theorem of Section 6 there exists a subsequence $\{u_{n_i}\}$ which converges uniformly to a (uniformly) continuous function v_0 on \mathbb{T} . We likewise define the iterates v_n of v_0 by $v_n = \mathfrak{F}v_{n-1} = \mathfrak{F}^n v_0$. If we convolve the members of the subsequence $\{u_{n_i}\}$ with F we obtain the next members of the original sequence and as $u_{n_i} \rightarrow v_0$ it follows that $u_{n_i+1} \rightarrow v_1$. Proceeding iteratively in this fashion, we obtain $u_{n_i+k} \rightarrow v_k$ uniformly for each k .

The functions u_n attain their maximum and minimum values. Let $M_n = \max u_n$ for each n . Now

$$u_{n+1}(t) = \int_{\mathbb{T}} u_n(t-x) dF(x) \leq M_n \int_{\mathbb{T}} dF(x) = M_n$$

for every t . Accordingly, $M_{n+1} \leq M_n$ and the sequence $\{M_n\}$ is decreasing. A similar argument shows that the sequence of minima of the functions u_n is increasing. Thus $\{M_n\}$ is a decreasing sequence bounded below by the minimum of u_0 ; it follows that $M_n \rightarrow M$ for some M . (Of course, a dual argument shows that the minima converge as well but we only need one of the extrema for the argument to follow.) As every subsequence of a convergent sequence converges to the same limit, it follows that $M_{n_i} \rightarrow M$, as well, so that this limiting value must perforce be the maximum of v_0 , $M = \max v_0$. But the same argument applied to the subsequence $\{u_{n_i+k}\}$ shows that $M_{n_i+k} \rightarrow M$ whence v_k has the same maximum as v_0 , $M = \max v_k$, for every k .

We've hence shown that the equicontinuous sequence of iterates v_0, v_1, \dots have a common maximum value M . We claim now that this is possible only if $v_0 = M$ is a constant. Suppose, to the contrary, that v_0 is strictly less than M at some point y . As v_0 is continuous, it follows that v_0 is strictly less than M everywhere in some interval \mathbb{I} containing y . If F is continuous then this immediately leads to a contradiction as $\int_{(t-\mathbb{I})} v_0(t-x) dF(x) < MF(t-\mathbb{I})$ where we use the notation $(t-\mathbb{I}) := \{(t-x) : x \in \mathbb{I}\}$ to represent an affine shift of \mathbb{I} . It follows that

$$v_1(t) = \int_{\mathbb{T}} v_0(t-x) dF(x) < MF(t-\mathbb{I}) + MF(\mathbb{T} \setminus (t-\mathbb{I})) = M$$

for every t and we have a contradiction. Suppose now that F has points of jump. As a single point of jump may trivially be placed at a vertex of an equilateral triangle, say, F has to have at least two distinct points of jump which cannot both be vertices of a regular polygon. As rotations do not affect maxima, we may suppose that F has one of these points of jump at 0, the other at ϑ in the real line modulo one. Then ϑ must be irrational, else a regular polygon can be found containing both 0 and ϑ as vertices. (If ϑ is rational and equal to a/b , say, a regular polygon with a vertex at 0 and side $1/b$ will serve.) It follows that $F * F$ has points of jump at 0, (ϑ) , and (2ϑ) . Proceeding by induction, F^{*k} will have points of jump at $0, (\vartheta), (2\vartheta), \dots, (k\vartheta)$. By Kronecker's theorem, the points $\{(k\vartheta), k \geq 1\}$ are dense in the unit interval. Consequently, for sufficiently large k and any t , the interval $(t-\mathbb{I})$ will contain at least one point x' which is one of the points of jump $0, (\vartheta), \dots, (k\vartheta)$. Then, $v_0(t-x') < M$, while $v_0(t-x) \leq M$ for $x \neq x'$. Suppose the jump of F at x' has size p . Then $0 < p < 1$ with p bounded above because there are at least two points of jump. It follows that

$$v_k(t) = \int_{\mathbb{T}} v_0(t-x) dF^{*k}(x) < M(1-p) + Mp = M,$$

a contradiction. It must hence be the case that v_0 takes the constant value M at all points of \mathbb{T} and all succeeding iterates v_k must then also be constant and equal to M . It follows that $u_{n_i+k} \rightarrow M$ for every k and, as all these subsequences converge to the same constant limit, the sequence $\{u_n\}$ itself converges uniformly to M .

Thus, for every continuous function u on \mathbb{T} , the iterates $\mathfrak{F}^n u(t)$ converge uniformly to a constant value. A minor adaptation of the proof of the equivalence theorem of Section 6 to continuous functions defined on the circle \mathbb{T} instead of the line shows that the sequence of distributions $\{F^{*n}\}$ converges vaguely to a limiting distribution H which must be proper because the circle is bounded. Thus, if \mathfrak{H} represents the operator corresponding to H , for every continuous u on the circle, we have $\mathfrak{F}^n u \rightarrow \mathfrak{H}u$ uniformly. It follows that the convolution with the limiting distribution, $\mathfrak{H}u(t) = \int_{\mathbb{T}} u(t-x) dH(x)$, is a constant for any continuous function u on \mathbb{T} . But this is the same situation encountered in the proof of Weyl's equidistribution theorem in the previous section. And our conclusion is the same: H must be the uniform distribution on the circle. ▶

Problems 19,20 examine what happens on the vertices of a regular polygon.

10 Problems

1. If $X_n \xrightarrow{d} X$ then $aX_n + b \xrightarrow{d} aX + b$ and if also $a_n \rightarrow a$ then $(a_n - a)X_n \xrightarrow{d} 0$. If $X_n \xrightarrow{d} X$, $a_n \rightarrow a$, and $b_n \rightarrow b$, then $a_n X_n + b_n \xrightarrow{d} aX + b$.
2. *Extension.* If $A_n \xrightarrow{d} a$, $B_n \xrightarrow{d} b$, and $X_n \xrightarrow{d} X$, then $A_n X_n + B_n \xrightarrow{d} aX + b$.
3. Suppose g is a continuous, real-valued function of a real variable. If $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$. The result holds even if only g is a Baire function, its set of discontinuities D_g is measurable, and $P\{X \in D_g\} = 0$.
4. Suppose X_1, X_2, \dots is a sequence of independent random variables with the common exponential distribution of mean $1/\alpha$. For each n , let $M_n = \max\{X_1, \dots, X_n\}$. Show that $M_n - \alpha^{-1} \log n$ converges in distribution and determine the limiting d.f.
5. *Convergent densities, convergent d.f.s.* Suppose X_n and X have densities f_n and f , respectively. If $f_n(x) \rightarrow f(x)$ for a.e. x , show that $X_n \xrightarrow{d} X$.
6. *Non-convergent densities, convergent d.f.s.* Let $f_n(x) = 1 + \sin 2\pi nx$ be a sequence of densities with support in the unit interval. Show that while the sequence $\{f_n\}$ does not converge, the corresponding sequence of d.f.s $\{F_n\}$ converges vaguely.
7. Let $\{X_n\}$ be an arbitrary sequence of random variables. Show that there exist positive constants a_n such that $a_n X_n \xrightarrow{d} 0$.
8. *Operator norm.* If \mathfrak{S} and \mathfrak{T} are bounded operators show that $\|\mathfrak{S} + \mathfrak{T}\| \leq \|\mathfrak{S}\| + \|\mathfrak{T}\|$ and $\|\mathfrak{ST}\| \leq \|\mathfrak{S}\| \cdot \|\mathfrak{T}\|$.
9. *Vague convergence.* If $F_n \xrightarrow{v} F$ and $G_n \xrightarrow{v} G$ show that $F_n * G_n \xrightarrow{v} F * G$.
10. *Once more, Lebesgue measure.* With the conventions of Section V.2, let $t = \sum_{k=1}^{\infty} z_k(t)2^{-k}$ be the binary expansion of each real number t in the unit interval $0 \leq t < 1$. Let A_n be the indicator of the set of points t for which $z_{n+1}(t) = \dots = z_{2n}(t) = 0$ and set $f_n(t) = 2^n 1_{A_n}(t)$. Show that $\{f_n\}$ is a sequence of densities converging pointwise to

zero except on a set of Lebesgue measure zero. On this exceptional set, redefine $f_n(t) = 0$ for all n , so that $f_n(t) \rightarrow 0$ everywhere. Show that the d.f.s corresponding to these densities converge vaguely to the uniform distribution on the unit interval.

11. Suppose F_n is concentrated at n^{-1} . Let $u(x) = \sin(x^2)$. Show that $\mathfrak{F}_n u \rightarrow u$ pointwise but *not* uniformly. Comment on what has failed in the equivalence theorem.

12. Inversion. Suppose the d.f. F has a moment μ_n of every order $n \geq 1$. If $\frac{1}{n} \mu_n^{1/n} \rightarrow 0$ then the sequence of moments $\{\mu_n, n \geq 1\}$ determines F . [Hint: Replace Laplace transforms by characteristic functions in the proof of Theorem XV.1.4.]

13. Sequences of characteristic functions. For each $n \geq 1$, let F_n be a probability distribution with characteristic function $\widehat{F}_n(\xi) = \int_{-\infty}^{\infty} e^{i\xi x} dF_n(x)$. Suppose there exists a function ψ such that $\widehat{F}_n(\xi) \rightarrow \psi(\xi)$ for all ξ . Show that there exists a (possibly defective) distribution F such that $F_n \xrightarrow{v} F$. [Hint: Couple Helly's selection principle with (4.2).]

14. Continuation. A continuous function which is the pointwise limit of a sequence of characteristic functions is itself a characteristic function. [Hint: F is defective if, and only if, the limit ψ is discontinuous at the origin.]

15. Triangular array. Consider the triangular array of Bernoulli trials $\{X_{n,k}, 0 \leq k \leq n, n \geq 1\}$ where, for each n , the variables $X_{n,0}, X_{n,1}, \dots, X_{n,n}$ comprising the n th row of the array are Bernoulli(α/\sqrt{n}) for a positive constant α (and n sufficiently large). For each n , consider the row sum $S_n = \sum_{k=0}^n X_{n,k} I\left(\frac{k}{\sqrt{n}}\right)$ where $I(x)$ is a suitably regular function on the positive half-line. Show that S_n has characteristic function $\widehat{F}_n(\xi) = E(e^{i\xi S_n}) = \prod_{k=1}^n \left[1 + \frac{\alpha}{\sqrt{n}} (e^{i\xi I(k/\sqrt{n})} - 1)\right]$. By passing to the limit show that $\widehat{F}_n(\xi)$ converges to the characteristic function $\widehat{F}(\xi)$ where $\log \widehat{F}(\xi) = \alpha \int_0^\infty (e^{i\xi I(x)} - 1) dx$.

16. Continuation, shot noise. Chance fluctuations in the arrival of electrons leads to the noise process called shot noise. It is assumed that electron arrivals are governed by a Poisson process of rate α and that each arriving electron produces a current whose intensity t seconds after arrival is $I(t)$. The shot noise current at time t is then the random variable $X(t) = \sum_{k=1}^{\infty} I(t - T_k)$ where T_k are the epochs of past electron arrivals. We may model the process by partitioning $(-\infty, t)$ into small subintervals with endpoints $t_k = t - k/\sqrt{n}$ for $0 \leq k \leq n$ with the variable $X_{n,k}$ of the previous problem connoting at most one arrival in the tiny interval $(t_{k+1}, t_k]$ (the definition of the Poisson process!). The row sums S_n hence approximate $X(t)$ and, by a passage to the limit, we see that the characteristic function of $X(t)$ is $\widehat{F}(\xi)$. By differentiating $\widehat{F}(\xi)$ prove that $E(X(t)) = \alpha \int_0^\infty I(x) dx$ and $\text{Var}(X(t)) = \alpha \int_0^\infty I(x)^2 dx$. This is *Campbell's theorem*.

17. Total variation distance. We may define the total variation distance (see Problem XVII.2) between two distributions F and G by $d_{TV}(F, G) = \sup \|\mathfrak{F}u - \mathfrak{G}u\|$ where the supremum is over all continuous functions u of unit (supremum) norm and vanishing at $\pm\infty$. Show that d_{TV} is a metric.

18. Continuation. If F and G have densities f and g , respectively, then $d_{TV}(F, G) = \frac{1}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| dx$. (It will suffice to prove the result for continuous f and g .)

19. Fractional parts of sums. If F is concentrated on the vertices of a regular polygon with one vertex at 0, show that F^{*n} converges to a uniform atomic distribution.

20. Continuation. Show that the preceding result need not hold if F does not have an atom at the origin.

Normal Approximation

The convergence of distributions has historically been the focus of much of the classical theory which is concerned with limit laws dealing with the asymptotic properties of distributions. Principal among these limit laws is the fabled central limit theorem: the convergence of the distribution of random sums to the normal distribution is certainly the most important, though by no means the only, example of convergence of a sequence of distributions to a limiting distribution. The material of this chapter constitutes a logical outgrowth of the ideas first encountered in a somewhat magical light in Chapter VI to settings of vast generality and in this chapter the reader will encounter several powerful versions of the principle of central tendency. As an additional bonus, the elementary method of proof of a quite general formulation of the result that is presented here helps expose the key features of the machinery; and the selected applications may both amuse and help to illustrate the theorem's rich history and character—and its unexpected reach into apparently mundane settings.

C 1, 5, 8
A 6, 7, 9–12
F 2–4

1 Identical distributions, the basic limit theorem

Operator methods have proved to be very fruitful in several areas of the advanced theory, notably in the theory of Markov processes, but their most striking success is perhaps in the ease and elegance with which they encompass the theory of normal approximation.

The reader has seen the normal distribution make a rather mysterious appearance in de Moivre's limit theorem for the binomial in Section VI.5 and in a limit theorem for order statistics (Problems IX.30,31). It transpires that these early results are special cases of a very general phenomenon that has come to be called central tendency.

The central limit theorem occupies pride of place among the pillars on which probability rests both by reasons of history and its abiding and continuing impact. The theorem itself has a long history; the first version was proved by Abraham de Moivre in 1733 with very many succeeding variants, particular

cases and refinements culminating in a general theorem of J. W. Lindeberg in 1922 which incorporated all the previous versions.¹ Following Lindeberg new results, extensions, and applications have continued to surface and the theorem in all its facets continues to swell in importance and application.

The original proofs of the central limit theorem were long and convoluted but have been gradually simplified to the point where they can be exposed by elementary methods. In this section we shall see how the operator methods of the previous section lend themselves to a strikingly simple proof of Lindeberg's theorem exhibited by H. F. Trotter in 1959.² Trotter's ideas have been persuasively aired by W. Feller in his seminal work³ but in spite of their intrinsic elegance and elementary nature appear to have been crowded out of the popular lore by the Fourier methods championed by P. Lévy that we have seen in Chapter VI. At the cost of a little repetition I have provided operator proofs of several versions of the central limit theorem in this chapter; the reader following along will be struck by how seamlessly the operator method adapts to different settings.

Remarkably, the only added element needed in Trotter's proof is Taylor's theorem from elementary calculus. We begin with a simple but important case which provides a model for the general result.

CENTRAL LIMIT THEOREM FOR IDENTICAL DISTRIBUTIONS Suppose X_1, X_2, \dots is a sequence of independent random variables drawn from a common distribution F with mean zero and variance one. For each n , let $S_n^* = (X_1 + \dots + X_n)/\sqrt{n}$. Then S_n^* converges in distribution to the standard normal; explicitly, for each $a < b$,

$$P\{a < S_n^* < b\} \rightarrow \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

If $E(X_k) = \mu$ and $\text{Var}(X_k) = \sigma^2$ then, writing $S_n = X_1 + \dots + X_n$ as usual, we may centre and scale the variables by replacing X_k by $(X_k - \mu)/\sigma$ and $S_n^* = S_n/\sqrt{n}$ by $S_n^* = (S_n - n\mu)/\sqrt{n}$. $\sigma = \sum_{k=1}^n (X_k - \mu)/\sigma$ and the theorem holds for the now properly normalised S_n^* . There is no loss in generality in assuming hence that the variables X_k have zero mean and unit variance.

EXAMPLES: 1) *Bernoulli trials.* If the X_k are Bernoulli trials with success probability p then $S_n^* = \frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - p)/\sqrt{pq} = (S_n - np)/\sqrt{npq}$ tends in distribution to the standard normal Φ . This is the theorem of de Moivre and Laplace that we proved by elementary methods in Section VI.5 (see also Problem VI.11).

¹J. W. Lindeberg, "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung", *Mathematische Zeitschrift*, vol. 15, pp. 211–235, 1922.

²H. F. Trotter, "An elementary proof of the central limit theorem", *Archiv der Mathematik*, vol. 9, pp. 226–234, 1959.

³W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, op. cit.

2) *Random walks.* Suppose each X_k takes the values ± 1 , each with probability $1/2$. Then X_k has mean zero and variance one. The quantity S_n represents a random walk on the line and S_n/\sqrt{n} converges in distribution to the standard normal. We had seen this as a corollary of de Moivre's theorem in Section VI.5.

3) *Waiting times.* Suppose X_k has the geometric distribution with atomic probabilities $w(k; p) = q^k p$ for positive integers k . Then S_n denotes the waiting time to the n th success and has the negative binomial distribution with corresponding atomic probabilities $w_n(k; p) = \binom{n}{k} (-q)^k p^n$ with mean nq/p and variance nq/p^2 . The central limit theorem tells us that the centred and scaled waiting time distribution corresponding to the normalised variable $S_n^* = (S_n - nq/p)/\sqrt{nq/p^2}$ converges vaguely to the standard normal.

4) *Uniform densities.* If X_k has the uniform density $u(x) = 1$ ($0 < x < 1$) then S_n has density given by the n -fold convolution $u_n = u^{*n}$ of u with itself. An explicit expression for u_n may be found in (IX.1.2), the corresponding d.f. U_n recovered by a formal term-by-term integration. As the normalised random variable $S_n^* = (S_n - n/2)/\sqrt{n/12}$ has d.f. U_n^* satisfying $U_n^*(x) = U_n(x\sqrt{n/12} + n/2)$, it follows via the central limit theorem that U_n^* converges vaguely to Φ .

5) *Exponential densities.* If X_k has exponential density $g(x; \alpha) = \alpha e^{-\alpha x}$ with support in the positive half-line, then S_n has the gamma density $g_n(x; \alpha) = \alpha^n x^{n-1} e^{-\alpha x}/n!$ ($x > 0$). As each X_k has mean $1/\alpha$ and variance $1/\alpha^2$, the centred and scaled gamma distribution with density $\sqrt{\frac{n}{\alpha^2}} g_n(x\sqrt{\frac{n}{\alpha^2}} + \frac{n}{\alpha}; \alpha)$ converges vaguely to Φ . ▶

Before diving into the proof, let us examine the setting with a view to identifying the key elements. Write F_n for the distribution of X_k/\sqrt{n} . Then $F_n(y) = F(\sqrt{n}y)$. As

$$S_n^* = \frac{1}{\sqrt{n}} X_1 + \cdots + \frac{1}{\sqrt{n}} X_n$$

is the sum of n independent variables with a common distribution F_n , the distribution of S_n^* is the n -fold convolution F_n^{*n} and the theorem may be succinctly expressed in the statement $F_n^{*n} \xrightarrow{v} \Phi$. As a starting point we should begin by characterising the behaviour of F_n .

It is clear that F_n is increasingly concentrated near the origin as n increases. Indeed, suppose that $\mathbb{I} = (a, b)$ is any given interval. Then $F_n(\mathbb{I}) = F(\sqrt{n}b) - F(\sqrt{n}a)$ and consequently $F_n(\mathbb{I})$ tends to zero if a and b are both strictly positive or if they are both strictly negative, and tends to one if a is strictly negative and b is strictly positive. Thus, $F_n(\mathbb{I}) \rightarrow 1$ if the origin is an interior point of \mathbb{I} and $F_n(\mathbb{I}) \rightarrow 0$ if the origin is an interior point of \mathbb{I}^c , or, in other words, the sequence of distributions $\{F_n\}$ converges vaguely to the Heaviside distribution H_0 concentrated at the origin. Equivalently, if \mathfrak{F}_n denotes the operator corresponding to the distribution F_n , then $\mathfrak{F}_n u \rightarrow H_0 u = u$ or, what is

the same thing, $\mathfrak{F}_n u - u \rightarrow 0$, uniformly for all $u \in \bar{\mathcal{C}}_b$. This is still too crude for our purposes as a sum over a large number of terms could conceivably stray far from the origin even if the summands are individually small.

The existence of the second moment allows us to get a handle on what the required rate of convergence is. As F_n has zero mean and variance $1/n$ we have $\int y^2 dF_n(y) = 1/n$ or, equivalently, $\int y^2 n dF_n(y) = 1$. This suggests that the “right” scaling to consider is $n\mathfrak{F}_n$ and we hence seek a uniform limit for $n(\mathfrak{F}_n u - u)$. As $\mathfrak{F}_n u(t) = \int u(t-y) dF_n(y)$, the nature of this limit is suggested by Taylor’s theorem. Assuming sufficient smoothness in u , a formal Taylor development of $u(t-y)$ around t gives the expansion $u(t-y) = u(t) - u'(t)y + \frac{1}{2}u''(t)y^2 - \dots$. If we integrate formally term-by-term, the first term yields $u(t)$, the second term yields a value 0 because F , hence F_n , has zero mean, and the quadratic term yields a value $\frac{1}{2n}u''(t)$ as F_n has variance $1/n$. Pushing forward boldly with this plan, we hence obtain the speculative approximation

$$\mathfrak{F}_n u(t) = \int_{-\infty}^{\infty} u(t-y) dF_n(y) \stackrel{?}{=} u(t) - 0 + \frac{1}{2n}u''(t) + \text{terms small compared to } \frac{1}{n}$$

which suggests that the rate of convergence of $\mathfrak{F}_n u$ to u is of the order of $1/n$. This useful observation sets up the key result that follows.

THE BASIC LEMMA I *Suppose u is a continuous function with three bounded derivatives. Then*

$$n\|\mathfrak{F}_n u - u - \frac{1}{2n}u''\| \rightarrow 0.$$

PROOF: Let u be any function with three bounded derivatives. We may then select M sufficiently large so that $|u''(x)|$ and $|u'''(x)|$ are both upper bounded by M for all x . We consider the function

$$q_t(y) = u(t-y) - u(t) + u'(t)y - \frac{1}{2}u''(t)y^2. \quad (1.1)$$

This choice, as we shall see, is dictated by our desire to exploit the fact that F has finite variance. By additivity of expectation, it is now easy to see that

$$\mathfrak{F}_n u(t) - u(t) - \frac{1}{2n}u''(t) = \int_{-\infty}^{\infty} q_t(y) dF_n(y) \quad (1.2)$$

as F_n has zero mean and variance $1/n$. We will need good estimates of $q_t(y)$ when y is small but, as F_n gets increasingly concentrated near the origin, can afford to be cavalier with estimates away from the origin. The Taylor development of $u(t-y)$ around t points the way.

Fix any $\epsilon > 0$ and consider the interval $\mathbb{I}_\epsilon = (-\epsilon, \epsilon)$. Then $F_n(\mathbb{I}_\epsilon) > 1 - \epsilon$ and $F_n(\mathbb{I}_\epsilon^c) < \epsilon$ for all sufficiently large n . Truncating the Taylor development of $u(t-y)$ around t to three terms shows that $q_t(y) = \frac{1}{6}u'''(\xi)y^3 = \frac{1}{6}u'''(\xi)y \cdot y^2$ for some ξ in the interval contained between $t-y$ and t . It follows that

$|q_t(y)| \leq \frac{1}{6}M\epsilon y^2 < M\epsilon y^2$ for all $y \in \mathbb{I}_\epsilon$ and all t . For $y \in \mathbb{I}_\epsilon^c$ we can afford to be sloppier. Truncating the Taylor development of $u(t-y)$ to two terms shows that $q_t(y) = \frac{1}{2}u''(\eta)y^2 - \frac{1}{2}u''(t)y^2$ for some η in the interval contained between $t-y$ and t . Consequently, $|q_t(y)| \leq My^2$ for all y and, in particular, for $y \in \mathbb{I}_\epsilon^c$, and for all t . We may now partition the domain of integration to obtain

$$\begin{aligned} |\mathfrak{F}_n u(t) - u(t) - \frac{1}{2n}u''(t)| &\leq \int_{\mathbb{I}_\epsilon} |q_t(y)| dF_n(y) + \int_{\mathbb{I}_\epsilon^c} |q_t(y)| dF_n(y) \\ &< M\epsilon \int_{\mathbb{I}_\epsilon} y^2 dF_n(y) + M \int_{\mathbb{I}_\epsilon^c} y^2 dF_n(y) = \frac{M\epsilon}{n} \int_{|x|<\epsilon\sqrt{n}} x^2 dF(x) + \frac{M}{n} \int_{|x|\geq\epsilon\sqrt{n}} x^2 dF(x), \end{aligned} \quad (1.3)$$

the final step following via the change of variable $x \leftarrow \sqrt{n}y$. As $\int x^2 dF(x) = 1$ is convergent, it follows that for all sufficiently large n the second integral on the right will be less than ϵ , while we may simply overbound the first integral by 1. It follows that there exists $N = N(\epsilon)$ independent of t so that

$$n|\mathfrak{F}_n u(t) - u(t) - \frac{1}{2n}u''(t)| < 2M\epsilon$$

for all $n \geq N$. As $\epsilon > 0$ may be chosen arbitrarily small, the claimed result follows. ▶

How do we leverage this result? Lindeberg's clever idea was to compare the distribution of S_n^* with that of a sum of independent normals. Some notation first. Suppose Z_1, Z_2, \dots are independent, standard normal random variables. For each n , form the sum $T_n^* = \frac{1}{\sqrt{n}} \sum_{k=1}^n Z_k$. Write Φ_n for the distribution of Z_k/\sqrt{n} . Then $\Phi_n(z) = \Phi(\sqrt{n}z)$ is the distribution of a normal with mean zero and variance $1/n$. The point of the whole idea is that *the normal distribution is closed under convolution*: T_n^* is normally distributed with mean zero and variance one, or, in other words, T_n^* has the standard normal distribution Φ . We can express this in notation by writing $\Phi_n^{*n} = \Phi$. Alternatively, if \mathfrak{N}_n and \mathfrak{N} are the operators corresponding to the normal distributions Φ_n and Φ , respectively, then $\mathfrak{N}_n^n = \mathfrak{N}$.

Lindeberg's replacement procedure consists of systematically replacing the variables X_k/\sqrt{n} by the normal variables Z_k/\sqrt{n} . In effect this compares \mathfrak{F}_n^n with \mathfrak{N}_n^n by comparing the operators \mathfrak{F}_n and \mathfrak{N}_n , one at a time. The idea is that the preceding lemma holds for any properly normalised distribution and in particular for \mathfrak{F}_n and \mathfrak{N}_n .

PROOF OF THE THEOREM: Let u be any function with three bounded derivatives. Then, as a consequence of the reductionist theorem of Section XIX.5, we have $\|\mathfrak{F}_n u - \mathfrak{N} u\| = \|\mathfrak{F}_n u - \mathfrak{N}_n^n u\| \leq n \|\mathfrak{F}_n u - \mathfrak{N}_n u\|$. Adding and subtracting the term $u + \frac{1}{2n}u''$ inside the function norm on the right, we obtain

$$\|\mathfrak{F}_n u - \mathfrak{N} u\| \leq n \|\mathfrak{F}_n u - u - \frac{1}{2n}u''\| + n \|\mathfrak{N}_n u - u - \frac{1}{2n}u''\|$$

via the triangle inequality. The preceding lemma now shows that both terms on the right tend to zero. As every smooth function *a fortiori* has three bounded derivatives, it follows that $\|\mathfrak{F}_n^n u - \mathfrak{N}u\| \rightarrow 0$ for all smooth functions, whence $F_n^{*n} \xrightarrow{v} \Phi$ by the equivalence theorem of Section XIX.5. ▶

Our proof shows that if F and G are any two centred distributions with unit variance then the operators \mathfrak{F}_n^n and \mathfrak{G}_n^n both converge to the same limiting operator, or, equivalently, F_n^{*n} and G_n^{*n} converge vaguely to the same limiting distribution. The choice of G as the standard normal distribution Φ is perspicacious as $\mathfrak{N}_n^n = \mathfrak{N}$, equivalently, $\Phi_n^{*n} = \Phi$, for all n . This identifies precisely what the limiting distribution is: $F_n^{*n} \xrightarrow{v} \Phi$ for all properly normalised distributions F with zero mean and unit variance.

The critical stability property $\Phi_n^{*n} = \Phi$ of the normal appears fortuitous; at its root, however, are two basic results: additivity of variance and the closure of the normal under convolutions. In slightly more graphic language, suppose Z is a standard normal random variable, Z_1, Z_2, \dots independent copies of Z . Then $Z_1 + \dots + Z_n$ has the same distribution as $\sqrt{n}Z$, which is just another way of saying $\Phi_n^{*n} = \Phi$. Implicit in our proof is the statement that the normal is the unique distribution with strictly positive variance for which this is true. The normal distribution is said to be *stable* in view of this property.⁴

The method of proof of the basic lemma adapts gracefully to other situations and, following a little detour which the reader may wish to defer for later reading, in Sections 5 and 8 we will consider two generalisations in different directions.



2 The value of a third moment

A limit law such as that of the previous section provides theoretical support for its use in a particular setting but to be really useful it should come equipped with an estimate of the error. These were developed independently by A. C. Berry and C-G. Esseen in 1941–1942 and are still the subject of current research.⁵ Naturally enough, we anticipate that the smoother the underlying distribution, the faster the convergence. The simplest

⁴The stability of the normal led P. Lévy in 1924 to initiate a study of a generalisation of this property. Suppose X is a random variable with distribution F (assumed centred without loss of generality), and X_1, X_2, \dots are independent copies of X . Then F is said to be *stable* if, for each n , $X_1 + \dots + X_n$ has the same distribution as $c_n X$ for some positive constant c_n . It can be shown that the norming constants have to be of the form $n^{1/\alpha}$ with $0 < \alpha \leq 2$; the value α is called the *characteristic exponent* of the distribution. Our proof of the central limit theorem shows that the normal is the unique stable distribution with strictly positive variance and characteristic exponent 2. Norming with characteristic exponents $\alpha < 2$ leads to other stable distributions and limit laws of a rather different character for variables without variance.

⁵A. C. Berry, "The accuracy of the Gaussian approximation to the sum of independent variates", *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941. C-G. Esseen, "On the Liapounov limit of error in the theory of probability", *Arkiv för matematik, astronomi och fysik*, vol. A28, pp. 1–19, 1942.

setting is that when the summands possess three moments. The importance of the result owes as much to its uncomplicated statement as it does to the elegance of the estimate.

THE BERRY–ESSEEN THEOREM Suppose X_1, X_2, \dots is a sequence of independent random variables drawn from a common distribution F with zero mean, unit variance, and bounded third moment $\int_{\mathbb{R}} |x|^3 dF(x) = \gamma$. Let $S_n = X_1 + \dots + X_n$ and write G_n for the distribution of S_n/\sqrt{n} . Then, for each n , we have $\sup_t |G_n(t) - \Phi(t)| \leq 3\gamma/\sqrt{n}$.

Hölder's inequality tells us that $E(|X|^3)^{1/3} \geq E(|X|^2)^{1/2}$ (see Problem XIV.36) and so $\gamma \geq 1$. The bound I have provided in the Berry–Esseen theorem is hence trite for $n \leq 9$. We may assume without loss of generality therefore that $n \geq 10$.

Write $D_n(t) = G_n(t) - \Phi(t)$ for the *discrepancy* between the distribution of the normalised sum S_n/\sqrt{n} and the standard normal. It is not obvious how a frontal attack on the discrepancy may be prosecuted but the convolutional smoothing theorem of Section XIX.3 suggests a profitable, if indirect, line of inquiry.

Suppose $K(x)$ is any d.f. For each $\tau > 0$, let $K_\tau(x) = K(\tau x)$. Then $\{K_\tau, \tau > 0\}$ is a family of d.f.s obtained by scaling, hence of the same type, which, as τ increases, gets increasingly concentrated at the origin; indeed, as $\tau \rightarrow \infty$, K_τ converges vaguely to the Heaviside distribution H_0 which places all its mass at the origin. Writing \mathcal{K}_τ for the convolutional operator associated with K_τ we see then that, as $\tau \rightarrow \infty$, $\mathcal{K}_\tau D_n(t) \rightarrow D_n(t)$ at points of continuity of D_n . As $\mathcal{K}_\tau D_n = K_\tau * G_n - K_\tau * \Phi$ is a smoothed, hence better behaved, version of D_n we are led to bound the discrepancy by considering how the smoothed version behaves. The key lies in the proper selection of the smoothing distribution (or *convolutional kernel*) K .

We select for the kernel K the absolutely continuous d.f. with density

$$\kappa(x) = \frac{2 \sin(x/2)^2}{\pi x^2} = \frac{1}{2\pi} \operatorname{sinc}\left(\frac{x}{2}\right)^2.$$

The function κ is called the *Fejér kernel*.⁶ It is clear that κ is positive, continuous, bounded, and integrable (as it decays like x^{-2}). It will become clear very shortly that κ is indeed properly normalised to unit area but the reader who wishes to reserve judgement should simply multiply κ by a norming constant to be determined.

Our first order of business—and this is where the bulk of the work will be expended—is to attempt to estimate the smoothed discrepancy $\mathcal{K}_\tau D_n(t) = \mathcal{K}_\tau G_n(t) - \mathcal{K}_\tau \Phi(t)$ where the d.f.s $\mathcal{K}_\tau G_n(t) = K_\tau * G_n(t)$ and $\mathcal{K}_\tau \Phi(t) = K_\tau * \Phi(t)$ are convolutionally smoothed versions of G_n and Φ , respectively. Convolutions suggest a passage to the Fourier transform and the reader should refresh her memory of the salient features of the transform by glancing at Sections VI.2,3. The road is eased by the development of a nonce notation and we pause to note an ancillary result which the reader has already seen in a different setting.

The characteristic function \widehat{K} associated with the d.f. K is given by

$$\widehat{K}(\xi) = \int_{-\infty}^{\infty} e^{i\xi x} dK(x) = \int_{-\infty}^{\infty} \kappa(x) e^{i\xi x} dx$$

⁶The Fejér kernel is fundamental in Fourier analysis where it arises for much the same reasons in the context of smoothing in a Cesàro mean.

where we identify the expression on the right with the ordinary Fourier transform of the Fejér kernel κ . The “hat” notation for transforms is getting a little overburdened and as we would still like to retain it for characteristic functions, we introduce the nonce notation $\mathcal{F}\kappa(\xi)$ for the integral expression on the right standing for the ordinary Fourier transform of κ . A frontal attack on the integral looks difficult but it already appears in a dissembling guise in Example VI.2.2. The unit triangular function $\Delta(x) = \max\{0, 1 - |x|\}$ has Fourier transform $\mathcal{F}\Delta(\xi) = \text{sinc}(\xi/2)^2$, both functions clearly continuous and integrable so that by the Fourier inversion theorem of Section VI.3,

$$\Delta(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{sinc}\left(\frac{\xi}{2}\right)^2 e^{-i\xi x} d\xi = \int_{-\infty}^{\infty} \kappa(\xi) e^{i\xi(-x)} d\xi.$$

By interchanging the rôles of x and ξ , we may by inspection identify the Fourier transform of κ with the unit triangular function and so

$$\widehat{\kappa}(\xi) = \mathcal{F}\kappa(\xi) = \Delta(-\xi) = \Delta(\xi) = \max\{0, 1 - |\xi|\}.$$

As we shall see, the key fact on which all else rests is that *the characteristic function $\widehat{\kappa}$ has support in a bounded interval*. [In passing, the sceptical reader may now verify that $\int_{-\infty}^{\infty} \kappa(x) dx = \widehat{\kappa}(0) = \Delta(0) = 1$ so that the normalisation of κ has been honest.]

One final observation and we will be ready to proceed.

LEMMA 1 *If F and G are any two d.f.s with respective characteristic functions \widehat{F} and \widehat{G} , the distribution $F \star G$ obtained by convolution has characteristic function $\widehat{F \star G} = \widehat{F} \cdot \widehat{G}$.*

PROOF: Problem XV.29 suggests the simplest demonstration: suppose X and Y are independent random variables with marginal d.f.s F and G , respectively. Then $X + Y$ has d.f. $F \star G$ whence $\widehat{F \star G}(\xi) = E(e^{i\xi(X+Y)}) = E(e^{i\xi X} e^{i\xi Y}) = E(e^{i\xi X}) E(e^{i\xi Y}) = \widehat{F}(\xi) \widehat{G}(\xi)$ as $e^{i\xi X}$ and $e^{i\xi Y}$ are independent, hence uncorrelated. ▶

Theorem XV.1.2 provides the corresponding result for Laplace transforms (with essentially the same proof) while Theorem VI.2.1(4) spells out the corresponding statement for the ordinary Fourier transform of the convolution of densities.

We are now equipped to return to the smoothed discrepancy $\mathfrak{K}_\tau D_n = \mathfrak{K}_\tau G_n - \mathfrak{K}_\tau \Phi$. The smoothed d.f.s $\widehat{\mathfrak{K}_\tau G_n} = K_\tau \star G_n$ and $\widehat{\mathfrak{K}_\tau \Phi} = K_\tau \star \Phi$ have characteristic functions $\widehat{\mathfrak{K}_\tau G_n}(\xi) = \widehat{K}_\tau(\xi) \widehat{G}_n(\xi)$ and $\widehat{\mathfrak{K}_\tau \Phi}(\xi) = \widehat{K}_\tau(\xi) \widehat{\Phi}(\xi)$, respectively, and we now only need to determine the individual characteristic functions. The scaling property of the Fourier transform [Theorem VI.2.1(3)] shows that the d.f. $K_\tau(x) = K(\tau x)$ has characteristic function $\widehat{K}_\tau(\xi) = \int_{\mathbb{R}} e^{i\xi x} dK_\tau(x) = \widehat{K}(\xi/\tau) = \Delta(\xi/\tau)$. As $S_n = X_1 + \dots + X_n$ is the sum of independent variables, by repeated applications of the previous lemma, the d.f. G_n of the normalised sum S_n/\sqrt{n} has characteristic function $\widehat{G}_n(\xi) = \widehat{F}(\xi/\sqrt{n})^n$ (see also Problem XV.30). Finally, the characteristic function of the standard normal d.f. Φ is the ordinary Fourier transform of the density ϕ and this is given in Example VI.2.4 to be $\widehat{\Phi}(\xi) = \mathcal{F}\phi(\xi) = e^{-\xi^2/2}$. It follows that $\widehat{\mathfrak{K}_\tau G_n}(\xi) = \Delta(\xi/\tau) \widehat{F}(\xi/\sqrt{n})^n$ and $\widehat{\mathfrak{K}_\tau \Phi}(\xi) = \Delta(\xi/\tau) e^{-\xi^2/2}$. As $\widehat{G}_n(\xi)$ and $\widehat{\Phi}(\xi)$ are characteristic functions, hence bounded absolutely by 1, the fact that $\widehat{K}_\tau(\xi) = \Delta(\xi/\tau)$ has bounded support carries the day: *the characteristic functions $\widehat{\mathfrak{K}_\tau G_n}(\xi)$ and $\widehat{\mathfrak{K}_\tau \Phi}(\xi)$ are both integrable*. It follows *vide* the

corollary to the inversion theorem of Section XIX.4 that the d.f.s $\mathfrak{K}_\tau G_n$ and $\mathfrak{K}_\tau \Phi$ have bounded continuous densities $g_{n,\tau} = (\mathfrak{K}_\tau G_n)'$ and $\phi_\tau = (\mathfrak{K}_\tau \Phi)'$, respectively, and these moreover are given by the Fourier inversion formulæ

$$g_{n,\tau}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Delta(\xi/\tau) \widehat{F}(\xi/\sqrt{n})^n e^{-i\xi t} d\xi,$$

$$\phi_\tau(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Delta(\xi/\tau) e^{-\xi^2/2} e^{-i\xi t} d\xi.$$

By subtraction of the two expressions and combining terms inside the integrals we obtain

$$g_{n,\tau}(t) - \phi_\tau(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Delta(\xi/\tau) [\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}] e^{-i\xi t} d\xi.$$

As the difference $g_{n,\tau}(t) - \phi_\tau(t)$ of the continuous densities $g_{n,\tau}$ and ϕ_τ is manifestly integrable, by another appeal to the Fourier inversion theorem we see that the difference $g_{n,\tau} - \phi_\tau$ has ordinary Fourier transform

$$\mathcal{F}(g_{n,\tau} - \phi_\tau)(\xi) = \Delta(\xi/\tau) [\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}].$$

The smoothed discrepancy $\mathfrak{K}_\tau D_n(t) = \mathfrak{K}_\tau G_n(t) - \mathfrak{K}_\tau \Phi(t)$ is the difference of d.f.s, hence vanishes at $\pm\infty$. As differentiation is additive, $(\mathfrak{K}_\tau D_n)'(t) = (\mathfrak{K}_\tau G_n)'(t) - (\mathfrak{K}_\tau \Phi)'(t) = g_{n,\tau}(t) - \phi_\tau(t)$, and so, by the derivative property of the Fourier transform [Theorem VI.2.1(5)], the ordinary Fourier transform of $\mathfrak{K}_\tau D_n(t)$ is hence given by

$$\mathcal{F}(\mathfrak{K}_\tau D_n)(\xi) = \Delta(\xi/\tau) \left[\frac{\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}}{-i\xi} \right].$$

Write $h_n(\xi)$ for the expression in square brackets on the right. The normalised sum S_n/\sqrt{n} has zero mean and unit variance and as $\widehat{F}(\xi/\sqrt{n})^n = \widehat{G}_n(\xi)$ is its characteristic function it follows that $\widehat{G}'_n(0) = 0$ and $\widehat{G}''_n(0) = 1/i^2 = -1$ (see Problem XV.31). By two applications of l'Hôpital's rule we see then that $h_n(0) = h'_n(0) = 0$ whence h_n is continuously differentiable at the origin, hence everywhere. As $\Delta(\xi/\tau)$ is continuous and has support in the bounded interval $[\tau, \tau]$ it follows that $\Delta(\xi/\tau)h_n(\xi)$ is continuous and integrable whence, by another appeal to the Fourier inversion theorem, we obtain

$$\mathfrak{K}_\tau D_n(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Delta(\xi/\tau) \left[\frac{\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}}{-i\xi} \right] e^{-i\xi t} d\xi.$$

And here, at last, the benefits of smoothing via the Fejér kernel are in view: the integral is only formally over an infinite domain as the spectrum of the kernel κ_τ has support only in the bounded interval $[-\tau, \tau]$. As $0 \leq \Delta(\xi/\tau) \leq 1$, the modulus inequality boils the estimate down to a purely analytical expression:

$$|\mathfrak{K}_\tau D_n(t)| \leq \frac{1}{2\pi} \int_{-\tau}^{\tau} \left| \frac{\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}}{\xi} \right| d\xi. \quad (2.1)$$

Two elementary facts simplify the computation of the integral on the right.

Normal Approximation

LEMMA 2 Suppose $z = it$ where t is real or $z = -t$ where t is positive. Then the inequality $|e^z - \sum_{k=0}^{n-1} z^k/k!| \leq |z|^n/n!$ holds for each positive integer n .

PROOF: I will provide the proof for the case $z = it$, the case when $z = -t$ being entirely similar. The remainder in Taylor's formula for e^{it} truncated to n terms is given by

$$R_n(t) = e^{it} - \sum_{k=0}^{n-1} \frac{(it)^k}{k!} = \frac{i^n}{(n-1)!} \int_0^t e^{i(t-x)} x^{n-1} dx.$$

By the modulus inequality again, we obtain $|R_n(t)| \leq \frac{1}{(n-1)!} \int_0^{|t|} x^{n-1} dx = |t|^n/n!$. ▶

LEMMA 3 If $|A| \leq C$ and $|B| \leq C$ then $|A^n - B^n| \leq nC^{n-1}|A - B|$.

PROOF: By factoring out $A - B$ from $A^n - B^n$ we obtain $A^n - B^n = (A - B)(A^{n-1} + A^{n-2}B + \dots + B^{n-1})$ and the claimed result follows immediately. ▶

In (2.1), identify $A = \hat{F}(\xi/\sqrt{n})$ and $B = e^{-\xi^2/(2n)}$. Then, by the triangle inequality, we have

$$|A - B| \leq |\hat{F}(\xi/\sqrt{n}) - 1 + \xi^2/(2n)| + |e^{-\xi^2/(2n)} - 1 + \xi^2/(2n)|.$$

By Lemma 2, we have $|e^{-t} - 1 + t| \leq t^2/2$ for all positive t . By identifying $t = \xi^2/(2n)$, we have

$$|e^{-\xi^2/(2n)} - 1 + \xi^2/(2n)| \leq \frac{(\xi^2/(2n))^2}{2} = \frac{\xi^4}{8n^2}.$$

Another application of Lemma 2 shows also that

$$\begin{aligned} |\hat{F}(\xi/\sqrt{n}) - 1 + \xi^2/(2n)| &= \left| \int_{-\infty}^{\infty} \left(e^{it\xi/\sqrt{n}} - 1 - \frac{it\xi}{\sqrt{n}} + \frac{t^2\xi^2}{2n} \right) dF(t) \right| \\ &\leq \int_{-\infty}^{\infty} \frac{|t\xi|^3}{6n^{3/2}} dF(t) = \frac{\gamma|\xi|^3}{6n^{3/2}} \end{aligned}$$

where $\gamma = \int_{\mathbb{R}} |t|^3 dF(t)$ is the absolute third moment of F . Our estimate also gives the bound

$$|\hat{F}(\xi/\sqrt{n})| \leq 1 - \xi^2/(2n) + \gamma|\xi|^3/(6n^{3/2}) \leq 1 - 5\xi^2/(18n) \quad (2.2)$$

provided that $|\xi| \leq 4\sqrt{n}/(3\gamma)$.⁷ We now have a candidate for the integral limits in (2.1): we simply set $\tau = 4\sqrt{n}/(3\gamma)$. The expression $e^{-5\xi^2/(18n)}$ dominates both $1 - 5\xi^2/(18n)$ and $e^{-\xi^2/(2n)}$ [the first because $1 - x \leq e^{-x}$, the second because $5/18 < 1/2$ (sic)]. We may hence select $C = e^{-5\xi^2/(18n)}$ in Lemma 3 and so, for $n \geq 10$, we have

$$C^{n-1} = \exp\left\{-\frac{5\xi^2}{18}(1 - \frac{1}{n})\right\} \leq e^{-\xi^2/4}.$$

⁷The peculiar choice of constant 5/18 in the bound (2.2) was dictated by hindsight. The reader who objects to strange constants being pulled out of thin air should simply replace 5/18 by a constant $0 < c < 1/2$ and, at the conclusion of the analysis, optimise over the choice of c .

Putting our estimates together, the integrand on the right in (2.1) is uniformly bounded by

$$\left| \frac{\widehat{F}(\xi/\sqrt{n})^n - e^{-\xi^2/2}}{\xi} \right| \leq nC^{n-1}|A - B| \leq \left(\frac{\gamma\xi^2}{6\sqrt{n}} + \frac{|\xi|^3}{8n} \right) e^{-\xi^4/4},$$

throughout the entire range of integration $|\xi| \leq \tau = 4\sqrt{n}/(3\gamma)$, the bound valid whenever $n \geq 10$. Replacing the integrand in (2.1) by the bound given above can only increase the value of the integral and now expanding the range of integration to the entire real line keeps the bound moving in the right direction. This leaves two elementary Gaussian moment integrals to evaluate and by performing the simple integrations we obtain

$$\begin{aligned} |\mathfrak{K}_\tau D_n(t)| &\leq \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\gamma}{6\sqrt{n}} \xi^2 e^{-\xi^2/4} d\xi + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{8n} |\xi|^3 e^{-\xi^2/4} d\xi \\ &= \frac{\gamma}{3\sqrt{\pi n}} + \frac{1}{\pi n} = \frac{4}{9\tau\sqrt{\pi}} + \frac{4}{3\gamma\tau\pi\sqrt{n}}, \end{aligned}$$

the bound on the right uniform in t . As $\gamma \geq 1$ and $n > 9$, we obtain a compact estimate of the smoothed discrepancy.

LEMMA 4 Let $\tau = 4\sqrt{n}/(3\gamma)$. Then, for all $n \geq 10$, we have

$$\sup_t |\mathfrak{K}_\tau D_n(t)| \leq \frac{4}{9\tau\sqrt{\pi}} + \frac{4}{9\tau\pi}.$$

It only remains to compare the discrepancy D_n with its smoothed version $\mathfrak{K}_\tau D_n$. An elementary argument will suffice.

LEMMA 5 For every $n \geq 1$ and $\tau > 0$, we have

$$\sup_t |D_n(t)| \leq 2 \sup_t |\mathfrak{K}_\tau D_n(t)| + \frac{12\sqrt{2}}{\pi^{3/2}\tau}.$$

PROOF: Write $\delta_n = \sup_t |D_n(t)|$ and $\delta_{n,\tau} = \sup_t |\mathfrak{K}_\tau D_n(t)|$. The discrepancy $D_n(t) = G_n(t) - \Phi(t)$ vanishes at $\pm\infty$ and has one-sided limits $D_n(t-)$ and $D_n(t+)$ everywhere. There hence exists t_0 such that $D_n(t_0-) = \delta_n$ or $D_n(t_0+) = \delta_n$. We may suppose that $D_n(t_0) = \delta_n$. Now G_n is a d.f., hence increasing, while $\Phi'(t) = \phi(t) \leq (2\pi)^{-1/2}$ and so the normal d.f. Φ has a bounded rate of growth. It follows that

$$D_n(t_0 + \lambda) = D_n(t_0) + (G_n(t_0 + \lambda) - G_n(t_0)) - (\Phi(t_0 + \lambda) - \Phi(t_0)) \geq \delta_n - \frac{\lambda}{\sqrt{2\pi}}$$

for all $\lambda > 0$. We may massage the left-hand side of the inequality into the form of a convolutional argument by setting $h = \delta_n\sqrt{\pi/2}$, $t_1 = t_0 + h$, and $y = h - \lambda$ to obtain $D_n(t_1 - y) \geq \delta_n - (h - y)/\sqrt{2\pi} = \delta_n/2 + y/\sqrt{2\pi}$ whenever $|y| \leq h$. When $|y| > h$ we may use the trivial bound $D_n(t_1 - y) \geq -\delta_n$. It follows that

$$\begin{aligned} \delta_{n,\tau} &\geq \mathfrak{K}_\tau D_n(t_1) = \int_{\mathbb{R}} D_n(t_1 - y) dK_\tau(y) \\ &\geq \int_{|y| \leq h} \left(\frac{\delta_n}{2} + \frac{y}{\sqrt{2\pi}} \right) dK_\tau(y) + \int_{|y| > h} (-\delta_n) dK_\tau(y). \end{aligned}$$

The term on the right with the linear integrand vanishes in view of the symmetry of the d.f. K_τ as its density $\kappa_\tau(y) = \tau\kappa(\tau y)$ is an even function: $\int_{|y| \leq h} y dK_\tau(y) = 0$. And, as K_τ is a proper distribution, we have $\int_{|y| \leq h} dK_\tau(y) = 1 - \int_{|y| > h} dK_\tau(y)$. By collecting terms, it follows that

$$\delta_{n,\tau} \geq \frac{\delta_n}{2} - \frac{3\delta_n}{2} \int_{|y| > h} dK_\tau(y).$$

To complete the proof we only need a tail estimate for the d.f. K_τ . Again elementary bounds serve. We have

$$\begin{aligned} \int_{|y| > h} dK_\tau(y) &= 2 \int_h^\infty \frac{\tau}{2\pi} \operatorname{sinc}\left(\frac{\tau y}{2}\right)^2 dy \stackrel{(u=\tau y/2)}{=} \frac{2}{\pi} \int_{\tau h/2}^\infty \operatorname{sinc}(u)^2 du \\ &\leq \frac{2}{\pi} \int_{\tau h/2}^\infty \frac{du}{u^2} = \frac{4}{\pi \tau h} = \frac{4\sqrt{2}}{\pi^{3/2} \tau \delta_n}. \end{aligned}$$

It follows that $\delta_{n,\tau} \geq \delta_n/2 - 6\sqrt{2}/(\pi^{3/2}\tau)$ as was to be shown. ▶

Putting the estimates from Lemmas 4 and 5 together, we obtain

$$\sup_t |D_n(t)| \leq \frac{8}{9\tau\pi} (\sqrt{\pi} + 1) + \frac{12\sqrt{2}}{\pi^{3/2}\tau} = \frac{3\gamma}{\sqrt{n}} \left[\frac{2}{9\pi} (\sqrt{\pi} + 1) + \frac{3\sqrt{2}}{\pi^{3/2}} \right].$$

The expression in square brackets on the right evaluates to approximately 0.958 which is less than one. This completes the proof of the Berry–Esseen theorem: *there exists a universal constant C such that $\sup_t |D_n(t)| \leq C\gamma/\sqrt{n}$.* Our proof shows that we may select $C = 3$.

The elegant proof I have provided is due to W. Feller.⁸ The constant "3" in the Berry–Esseen bound can be improved, though by increasingly arduous calculations: circa 2011 the best universal constant for the Berry–Esseen bound is approximately $C < 0.48$.⁹ There is not much room for further improvement as Esseen showed as early as 1956 that any such universal constant C must be at least $(\sqrt{10} + 3)/(6\sqrt{2\pi}) \approx 0.41$.¹⁰



3 Stein's method

The classical Fourier analytical proofs of the Berry–Esseen theorem are of a wondrous delicacy and provide very precise results as we saw in the previous section. The price to be paid is in the heavy machinery of Fourier theory that is requisite. Of much more recent vintage is the subtle approximation method of C. Stein which provides an alternative argument, remarkably elementary, at the expense of a little precision. This takes

⁸W. Feller, "On the Berry–Esseen theorem", *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 10, pp. 261–268, 1968.

⁹V. Yu. Korolev and I. G. Shevtsova, "On the upper bound for the absolute constant in the Berry–Esseen inequality", *Theory of Probability and its Applications*, vol. 54, no. 4, pp. 638–658, 2010.

¹⁰C-G. Esseen, "A moment inequality with an application to the central limit theorem", *Skand. Aktuarietidskr.*, vol. 39, pp. 160–170, 1956.

us a little afield from the theme of convolutional operators but complements the Stein-Chen method of Poisson approximation so beautifully that I cannot resist providing the reader with a vignette of how Stein's basic idea is adapted to normal approximation. This section and the next are independent of the rest of this chapter and are perhaps best appreciated in conjunction with the material of Chapter XVIII.

We start with some notation. For each real t , write $h_t(x) = 1_{(-\infty, t]}(x)$ and, for each $\lambda > 0$, introduce the linear interpolation

$$h_{t,\lambda}(x) = \begin{cases} 1 & \text{if } x \leq t, \\ 1 - (x - t)/\lambda & \text{if } t < x \leq t + \lambda, \\ 0 & \text{if } x > t + \lambda. \end{cases}$$

If h is a Baire function it is customary to abuse notation and, with the usual caveat on existence, write $\Phi(h) := \int_{-\infty}^{\infty} h(x)\phi(x) dx$ for the expectation of h with respect to the standard normal distribution $N(0, 1)$. When h is the step function h_t we recover the d.f. of the normal, $\Phi(h_t) = \Phi(1_{(-\infty, t]}) = \Phi(t)$. When h is selected to be the linear interpolation $h_{t,\lambda}$ we also write $\Phi(h_{t,\lambda}) = \Phi_{\lambda}(t)$. This abuse of notation is harmless and will be restricted to this section and the next.

Begin with a random variable X with d.f. F . The discrepancy $D(t) = F(t) - \Phi(t)$ tells us how well the normal approximates F . Writing the discrepancy in the evocative form $D(t) = E(h_t(X) - \Phi(t))$ suggests a pathway to estimating it via the expectation of smooth functions. The function h_t itself is awkward to work with because of the discontinuity at t and we replace it by $h_{t,\lambda}$ leading to the smoothed discrepancy given by $D_{\lambda}(t) = E(h_{t,\lambda}(X) - \Phi_{\lambda}(t))$. Naturally enough, we expect the behaviour of $D(t)$ to track that of $D_{\lambda}(t)$ when λ is small and it will be convenient to establish this first.

LEMMA 1 *The inequality $\sup_t |D(t)| \leq \sup_t |D_{\lambda}(t)| + \lambda/\sqrt{2\pi}$ holds for each $\lambda > 0$.*

PROOF: The proof rests upon the domination condition $h_{t-\lambda,\lambda}(x) \leq h_t(x) \leq h_{t,\lambda}(x)$. This leads to the twin inequalities

$$\begin{aligned} (h_{t-\lambda,\lambda}(X) - \Phi_{\lambda}(t - \lambda)) - (\Phi(t) - \Phi_{\lambda}(t - \lambda)) &\leq h_t(X) - \Phi(t) \\ &\leq (h_{t,\lambda}(X) - \Phi_{\lambda}(t)) + (\Phi_{\lambda}(t) - \Phi(t)). \end{aligned} \quad (3.1)$$

As $0 < \Phi'(x) = \phi(x) \leq (2\pi)^{-1/2}$, the normal d.f. has a bounded rate of growth and accordingly

$$\begin{aligned} 0 &\leq \Phi(t) - \Phi_{\lambda}(t - \lambda) \leq \Phi(t) - \Phi(t - \lambda) \leq \lambda/\sqrt{2\pi}, \\ 0 &\leq \Phi_{\lambda}(t) - \Phi(t) \leq \Phi(t + \lambda) - \Phi(t) \leq \lambda/\sqrt{2\pi}. \end{aligned} \quad (3.2)$$

By taking expectations, the upper bound of (3.1) shows hence that $D(t) \leq D_{\lambda}(t) + \lambda/\sqrt{2\pi}$ while the lower bound of (3.1) shows that $-D(t) \leq -D_{\lambda}(t - \lambda) + \lambda/\sqrt{2\pi}$. Taking supremums of both sides of these inequalities establishes the claimed result. ▶

The essence of Stein's method is to replace the smoothed discrepancy $D_{\lambda}(t) = E(h_{t,\lambda}(X) - \Phi_{\lambda}(t))$ by the difference of even better behaved functions. A clue may be gleaned by taking expectations with respect to the normal. Suppose, to begin, that

Normal Approximation

X is a standard normal random variable. If g is a suitably smooth function satisfying $\lim g(x)\phi(x) = 0$ as $x \rightarrow \pm\infty$ then an easy integration by parts shows that

$$\begin{aligned} \mathbb{E}(g'(X)) &= \int_{-\infty}^{\infty} g'(x)\phi(x) dx = g(x)\phi(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} g(x)\phi'(x) dx \\ &= \int_{-\infty}^{\infty} xg(x)\phi(x) dx = \mathbb{E}(Xg(X)), \end{aligned}$$

or, $\mathbb{E}(g'(X) - Xg(X)) = 0 = \mathbb{E}(h_{t,\lambda}(X) - \Phi_\lambda(t))$, the latter equality holding by definition for each t when $X \sim \mathcal{N}(0, 1)$. This is suggestive. We are now led to consider solutions of *Stein's equation for the normal*

$$g'(x) - xg(x) = h(x) - \Phi(h) \quad (3.3)$$

for suitably regular Baire functions h . This is the analogue for the normal of Stein's equation for the Poisson that the reader had encountered in (XVIII.1.7). We now seek an analogue of (XVIII.1.8) for an explicit solution of (3.3).

It will be convenient to introduce a nonce terminology. We say that a real-valued function g on the real line is *admissible* if it is continuously differentiable and $g(x)$, $xg(x)$, and $g'(x)$ are all absolutely bounded.

LEMMA 2 *Suppose h is bounded and continuous. Then Stein's equation (3.3) has an admissible solution g .*

PROOF: Setting $f(x) = h(x) - \Phi(h)$, we may rewrite Stein's equation (3.3) in the form $\frac{d}{dx}(g(x)\phi(x)) = f(x)\phi(x)$. This equation has a unique solution up to a constant which we may as well set to zero. Integrating out we see that

$$g(x) = \frac{1}{\phi(x)} \int_{-\infty}^x f(y)\phi(y) dy \quad (3.4)$$

is a continuously differentiable solution of (3.3). [This is the analogue of (XVIII.1.8) and is the only bounded solution; all other solutions differ from this one by the addition of a term of the form $c/\phi(x)$ for a constant c .] As $\int_{-\infty}^{\infty} (h(y) - \Phi(h))\phi(y) dy = 0$, we may write (3.4) in the equivalent forms

$$g(x) = \frac{1}{\phi(x)} \int_{-\infty}^x (h(y) - \Phi(h))\phi(y) dy = \frac{-1}{\phi(x)} \int_x^{\infty} (h(y) - \Phi(h))\phi(y) dy. \quad (3.4')$$

Suppose now that $\sup_x |h(x)| \leq C$. Then, by the modulus inequality applied to each of the integral forms in (3.4'), we obtain $|g(x)| \leq 2C \min\{\Phi(x), 1 - \Phi(x)\}/\phi(x)$. The normal tail bounds of Lemma VI.1.3 and Theorem X.1.1 show that $\min\{\Phi(x), 1 - \Phi(x)\} \leq \min\{\sqrt{\frac{\pi}{2}}, \frac{1}{x}\}\phi(x)$. The first of the bounds on the right shows that $|g(x)| \leq \sqrt{2\pi}C$, while the second shows that $|xg(x)| \leq 2C$. And, finally, we obtain from (3.3) via the triangle inequality that $|g'(x)| \leq |h(x)| + |\Phi(h)| + |xg(x)| \leq 4C$. ▶

We are now ready for Stein's characterisation of the normal; this is the normal analogue of Theorem XVIII.1.2.

THEOREM A random variable X has a standard normal distribution if, and only if,

$$E(g'(X) - Xg(X)) = 0$$

for every admissible g .

PROOF: Necessity has already been demonstrated. To prove sufficiency fix t and λ and take for g the solution to Stein's equation (3.3) with $h = h_{t,\lambda}$. Then $E(h_{t,\lambda}(X) - \Phi_\lambda(t)) = E(g'(X) - Xg(X)) = 0$ or, equivalently, $E(h_{t,\lambda}(X)) = \Phi_\lambda(t)$. The random variables $h_{t,\lambda}(X)$ occurring on the left of the identity are bounded (trivially by 1) and converge pointwise to the step function $h_t(X)$ as $\lambda \rightarrow 0$. By the dominated convergence theorem it follows that $E(h_{t,\lambda}(X)) \rightarrow E(h_t(X)) = P\{X \leq t\}$. On the other hand, in view of (3.2), $\Phi_\lambda(t)$ converges to $\Phi(t)$ as $\lambda \rightarrow 0$. It follows that the limits coincide and so $P\{X \leq t\} = \Phi(t)$ and we see that X has a standard normal distribution. ▶

Our theorem suggests an analogue of the Poisson approximation slogan of Section XVIII.2, this time for normal approximation.

SLOGAN A random variable X is approximately standard normal if $E(g'(X) - Xg(X)) \approx 0$ for a suitable selection of admissible g .

In detail, for each t , we select g to be a solution of Stein's equation (3.3) for $h = h_{t,\lambda}$ and attempt to show that $E(g'(X) - Xg(X)) = E(h_{t,\lambda}(X) - \Phi_\lambda(t))$ is uniformly small in t for a suitably small selection of λ . In the following section we will put the idea to work in the context of the rate of convergence in the central limit theorem.

4 Berry–Esseen revisited

We deal with a sequence X_1, X_2, \dots of independent random variables drawn from a common distribution F with mean zero, variance one, and a finite absolute third moment $\gamma = \int_{\mathbb{R}} |x|^3 dF(x)$. For each n , we write $S_n = X_1 + \dots + X_n$ and let G_n denote the d.f. of the normalised sum S_n/\sqrt{n} . Introducing subscripts to keep the rôle of n in view, write $D_n(t) = G_n(t) - \Phi(t) = E[h_t(S_n/\sqrt{n}) - \Phi(t)]$ and $\delta_n = \sup_t |D_n(t)|$ for the discrepancy, respectively, maximal discrepancy, and likewise $D_{n,\lambda}(t) = E[h_{t,\lambda}(S_n/\sqrt{n}) - \Phi_\lambda(t)]$ and $\delta_{n,\lambda} = \sup_t |D_{n,\lambda}(t)|$ for the smoothed discrepancy.

In view of Lemma 3.1, $\delta_n \leq \delta_{n,\lambda} + \lambda/\sqrt{2\pi}$ and we may focus on the smoothed discrepancy. Let g be the solution (3.4) of Stein's equation for the choice $h = h_{t,\lambda}$. Then

$$\begin{aligned} D_{n,\lambda}(t) &= E\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{S_n}{\sqrt{n}}g\left(\frac{S_n}{\sqrt{n}}\right)\right] = E\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \frac{1}{\sqrt{n}} \sum_{j=1}^n X_j g\left(\frac{S_n}{\sqrt{n}}\right)\right] \\ &= E\left[g'\left(\frac{S_n}{\sqrt{n}}\right) - \sqrt{n} X_n g\left(\frac{S_n}{\sqrt{n}}\right)\right]. \quad (4.1) \end{aligned}$$

The final step follows because the variables X_1, \dots, X_n are exchangeable so that the pairs $(X_1, S_n), \dots, (X_n, S_n)$ all have the same distribution whence, by additivity of expectation,

$$E\left[\sum_{j=1}^n X_j g\left(\frac{S_n}{\sqrt{n}}\right)\right] = \sum_{j=1}^n E\left[X_j g\left(\frac{S_n}{\sqrt{n}}\right)\right] = n E\left[X_n g\left(\frac{S_n}{\sqrt{n}}\right)\right].$$

Normal Approximation

Now here's the clever bit. We attempt to evaluate the right-hand side of (4.1) recursively. Writing $S_n = S_{n-1} + X_n$, by the fundamental theorem of calculus, we have

$$g\left(\frac{S_{n-1}}{\sqrt{n}} + \frac{X_n}{\sqrt{n}}\right) - g\left(\frac{S_{n-1}}{\sqrt{n}}\right) = \int_{\frac{S_{n-1}}{\sqrt{n}}}^{\frac{S_{n-1}}{\sqrt{n}} + \frac{X_n}{\sqrt{n}}} g'(u) du = \frac{X_n}{\sqrt{n}} \int_0^1 g'\left(\frac{S_{n-1}}{\sqrt{n}} + v \frac{X_n}{\sqrt{n}}\right) dv$$

by a simple change of variable inside the integral. By multiplying both sides by $\sqrt{n} X_n$ and taking expectations, it follows that

$$\begin{aligned} \mathbb{E}\left[\sqrt{n} X_n g\left(\frac{S_n}{\sqrt{n}}\right)\right] &= \mathbb{E}\left[X_n^2 \int_0^1 g'\left(\frac{S_{n-1}}{\sqrt{n}} + v \frac{X_n}{\sqrt{n}}\right) dv\right] + \sqrt{n} \mathbb{E}\left[X_n g\left(\frac{S_{n-1}}{\sqrt{n}}\right)\right] \\ &= \mathbb{E}\left[X_n^2 \int_0^1 \left\{g'\left(\frac{S_{n-1}}{\sqrt{n}} + v \frac{X_n}{\sqrt{n}}\right) - g'\left(\frac{S_{n-1}}{\sqrt{n}}\right)\right\} dv\right] \\ &\quad + \mathbb{E}\left[X_n^2 g'\left(\frac{S_{n-1}}{\sqrt{n}}\right)\right] + \sqrt{n} \mathbb{E}\left[X_n g\left(\frac{S_{n-1}}{\sqrt{n}}\right)\right]. \end{aligned}$$

Now X_n and S_{n-1} are independent, *a fortiori* uncorrelated. The penultimate expectation on the right is hence given by $\mathbb{E}(X_n^2) \mathbb{E}[g'(S_{n-1}/\sqrt{n})] = \mathbb{E}[g'(S_{n-1}/\sqrt{n})]$ while the final expectation on the right is given by $\mathbb{E}(X_n) \mathbb{E}[g(S_{n-1}/\sqrt{n})] = 0$. (The reader should bear in mind that X_n has zero mean and unit variance.) Introducing the nonce notation

$$\Delta g'_n(v) = g'\left(\frac{S_{n-1}}{\sqrt{n}} + v \frac{X_n}{\sqrt{n}}\right) - g'\left(\frac{S_{n-1}}{\sqrt{n}}\right),$$

for $0 \leq v \leq 1$, we hence obtain

$$\mathbb{E}\left[\sqrt{n} X_n g\left(\frac{S_n}{\sqrt{n}}\right)\right] = \mathbb{E}\left[X_n^2 \int_0^1 \Delta g'_n(v) dv\right] + \mathbb{E}\left[g'\left(\frac{S_{n-1}}{\sqrt{n}}\right)\right].$$

By substitution back in (4.1) and collecting terms we obtain the compact expression

$$D_{n,\lambda}(t) = \mathbb{E}[\Delta g'_n(1)] - \int_0^1 \mathbb{E}[X_n^2 \Delta g'_n(v)] dv \tag{4.2}$$

and our problem simplifies to that of finding good estimates for the derivative differences $\Delta g'_n(v)$.

With $h = h_{t,\lambda}$ in Stein's equation (3.3), we have

$$g'(x+y) - g'(x) = yg(x+y) + x(g(x+y) - g(x)) + (h_{t,\lambda}(x+y) - h_{t,\lambda}(x)).$$

We can write the latter two terms on the right in a slightly more informative fashion by observing, first, that by the mean value theorem there exists $z = z(x, y)$ between x and $x+y$ such that $g(x+y) - g(x) = yg'(z)$ and, second, that $h'_{t,\lambda}(x) = -1_{(t,t+\lambda]}(x)/\lambda$ (excepting only at the discrete points $x = t$ and $x = t+\lambda$) so that by the fundamental theorem of calculus,

$$h_{t,\lambda}(x+y) - h_{t,\lambda}(x) = \int_x^{x+y} h'_{t,\lambda}(u) du = \frac{-y}{\lambda} \int_0^1 1_{(t,t+\lambda]}(x+wy) dw.$$

The final step follows by the simple change of variable $u = x + wy$ inside the integral. As $h_{t,\lambda}$ is bounded between 0 and 1, so is its Gaussian expectation $\Phi_\lambda(t)$, and hence $|h_{t,\lambda}(x) - \Phi_\lambda(t)| \leq 1$. We may hence improve on the bounds obtained at the end of the proof of Lemma 3.2 slightly and obtain $|g(x)| \leq \sqrt{\pi/2}$, $|xg(x)| \leq 1$, and $|g'(x)| \leq 2$. It follows by two applications of the triangle inequality that

$$|g'(x+y) - g'(x)| \leq |y| \left[\sqrt{\frac{\pi}{2}} + 2|x| + \frac{1}{\lambda} \int_0^1 1_{(t,t+\lambda]}(x+wy) dw \right].$$

Specialising by identifying x with S_{n-1}/\sqrt{n} and y with vX_n/\sqrt{n} (where $0 \leq v \leq 1$), we obtain

$$|\Delta g'_n(v)| \leq \frac{1}{\sqrt{n}} \left[\sqrt{\frac{\pi}{2}} |X_n| + 2|X_n| \cdot \frac{|S_{n-1}|}{\sqrt{n}} + \frac{1}{\lambda} \int_0^1 |X_n| \cdot 1_{(t,t+\lambda]} \left(\frac{S_{n-1}}{\sqrt{n}} + vw \frac{X_n}{\sqrt{n}} \right) dw \right]. \quad (4.3)$$

We now take expectations of both sides, the computation of the expectations of the terms on the right facilitated by simple observations.

We recall, to begin, that $E(X_n) = 0$, $E(X_n^2) = 1$, and $E(|X_n|^3) = \gamma \geq 1$, the last by Hölder's inequality. By the Cauchy–Schwarz inequality we may hence bound

$$\begin{aligned} E(|X_n|) &\leq E(X_n^2)^{1/2} = 1, \\ E(|S_{n-1}|/\sqrt{n}) &\leq E(|S_{n-1}|/\sqrt{n-1}) \leq E(S_{n-1}^2/(n-1))^{1/2} = 1. \end{aligned}$$

As X_n and S_{n-1} are independent, moreover, we have

$$E(|X_n| \cdot |S_{n-1}|/\sqrt{n}) = E(|X_n|) E(|S_{n-1}|/\sqrt{n}) \leq 1.$$

Finally, for all real ξ , we have

$$E \left[1_{(t,t+\lambda]} \left(\frac{S_{n-1}}{\sqrt{n}} + \xi \right) \right] = P \left\{ \sqrt{\frac{n}{n-1}} (t - \xi) < \frac{S_{n-1}}{\sqrt{n-1}} \leq \sqrt{\frac{n}{n-1}} (t - \xi + \lambda) \right\}.$$

Writing $a = \sqrt{\frac{n}{n-1}} (t - \xi)$ and $b = \sqrt{\frac{n}{n-1}} (t - \xi + \lambda)$, we see that

$$\begin{aligned} E \left[1_{(t,t+\lambda]} \left(\frac{S_{n-1}}{\sqrt{n}} + \xi \right) \right] &= E \left[h_b \left(\frac{S_{n-1}}{\sqrt{n-1}} \right) - h_a \left(\frac{S_{n-1}}{\sqrt{n-1}} \right) \right] \\ &= E[D_{n-1}(b)] - E[D_{n-1}(a)] + \Phi(b) - \Phi(a). \end{aligned}$$

Here's the crux: the expression on the right involves the discrepancy in one fewer dimension! As $\phi(x) \leq (2\pi)^{-1/2}$ we have $\Phi(b) - \Phi(a) \leq (b - a)/\sqrt{2\pi}$ and so, by taking absolute values, we see hence that

$$0 \leq E \left[1_{(t,t+\lambda]} \left(\frac{S_{n-1}}{\sqrt{n}} + \xi \right) \right] \leq 2\delta_{n-1} + \sqrt{\frac{n}{n-1}} \frac{\lambda}{\sqrt{2\pi}} \leq 2\delta_{n-1} + \frac{\lambda}{\sqrt{\pi}}$$

for all ξ and all $n \geq 2$. To compute the expectation of the integrand on the right in (4.3), by Fubini's theorem, we first condition on X_n and compute the expectation integral over S_{n-1} (identify vwX_n/\sqrt{n} with ξ) before finishing up by integrating over X_n to get

$$E \left[|X_n| \cdot 1_{(t,t+\lambda]} \left(\frac{S_{n-1}}{\sqrt{n}} + vw \frac{X_n}{\sqrt{n}} \right) \right] \leq \left(2\delta_{n-1} + \frac{\lambda}{\sqrt{\pi}} \right) E(|X_n|) \leq 2\delta_{n-1} + \frac{\lambda}{\sqrt{\pi}}.$$

Normal Approximation

The final step follows again courtesy the Cauchy–Schwarz inequality as $E(|X_n| \cdot 1) \leq E(X_n^2)^{1/2} = 1$. The pieces are all in place. By taking expectation of both sides of (4.3) we obtain

$$E(|\Delta g'_n(v)|) \leq \frac{2\delta_{n-1}}{\lambda\sqrt{n}} + \frac{1}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}} \right) \quad (4.4)$$

for all $0 \leq v \leq 1$ and *a fortiori* for $v = 1$. An entirely similar argument serves to estimate $E(|X_n^2 \Delta g'_n(v)|)$: all that is required is to replace $|X_n|$ throughout by $|X_n|^3$ in the argument above with little more needed than to replace stray references to $E(|X_n|)$ by the absolute third moment $E(|X_n|^3) = \gamma$. We hence obtain

$$E(|X_n^2 \Delta g'_n(v)|) \leq \frac{2\gamma\delta_{n-1}}{\lambda\sqrt{n}} + \frac{\gamma}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}} \right) \quad (4.5)$$

for all $0 \leq v \leq 1$. Returning to (4.2) with (4.4) and (4.5) in hand, we may bound the smoothed discrepancy by

$$|D_{n,\lambda}(t)| \leq \frac{2(1+\gamma)\delta_{n-1}}{\lambda\sqrt{n}} + \frac{1+\gamma}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}} \right) \leq \frac{4\gamma\delta_{n-1}}{\lambda\sqrt{n}} + \frac{2\gamma}{\sqrt{n}} \left(\sqrt{\frac{\pi}{2}} + 2 + \frac{1}{\sqrt{\pi}} \right),$$

the observation $1 + \gamma \leq 2\gamma$ [Hölder's inequality $E(|X|^3)^{1/3} \geq E(|X|^2)^{1/2}$, again] simplifying the bound further at the expense of a little precision. As the right-hand side is uniform in t , it also bounds $\delta_{n,\lambda} = \sup_t |D_{n,\lambda}(t)|$ and as, by Lemma 3.1, $\delta_n \leq \delta_{n,\lambda} + \lambda/\sqrt{2\pi}$, it is clear that we should select λ to decay like $n^{-1/2}$ to match the rates of decay of the various terms. We may as well select λ to minimise the algebra and by setting $\lambda = 8\gamma/\sqrt{n}$ we obtain the beautiful recurrence

$$\delta_n \leq \frac{\delta_{n-1}}{2} + \frac{c\gamma}{\sqrt{n}}$$

where $c = \sqrt{2\pi} + 4 + 2(1 + 2\sqrt{2})/\sqrt{\pi}$. By induction, it follows that

$$\delta_n \leq c\gamma \left(\frac{1}{\sqrt{n}} + \frac{1}{2\sqrt{n-1}} + \frac{1}{4\sqrt{n-2}} + \cdots + \frac{1}{2^{n-1}\sqrt{n-(n-1)}} \right). \quad (4.6)$$

The ratio of successive terms of the series $\sum_{k=0}^{n-1} 1/(2^k \sqrt{n-k})$ is bounded by

$$\frac{1}{2^{k+1}\sqrt{n-k-1}} / \frac{1}{2^k\sqrt{n-k}} = \frac{1}{2} \sqrt{\frac{n-k}{n-k-1}} = \frac{1}{2} \sqrt{1 + \frac{1}{n-k-1}} \leq \frac{1}{2} \sqrt{2} = \frac{1}{\sqrt{2}},$$

and as the bound on the right is less than one the terms of the series are decreasing. It follows that we may bound the sum inside the round brackets in (4.6) by a geometric series

$$\sum_{k=0}^{n-1} \frac{1}{2^k \sqrt{n-k}} \leq \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} \left(\frac{1}{\sqrt{2}} \right)^k \leq \frac{1}{\sqrt{n}} \sum_{k=0}^{\infty} \left(\frac{1}{\sqrt{2}} \right)^k = \frac{1}{\sqrt{n} (1 - 1/\sqrt{2})}.$$

We've hence unearthed anew the Berry–Esseen theorem: *there exists an absolute positive constant C such that $\delta_n \leq C\gamma/\sqrt{n}$* . Our demonstration shows that we may select $C =$

$[\sqrt{2\pi} + 4 + 2(1+2\sqrt{2})/\sqrt{\pi}] / (1 - 1/\sqrt{2}) < 37$. The constant is not particularly sharp—Section 2 of this chapter shows that we may select it to be 3—but the subtlety attained by such an elementary method of proof is impressive and it has the great virtue of being extensible to other situations. This proof of a slightly weakened form of the Berry–Esseen theorem is due to E. Bolthausen.¹¹

5 Varying distributions, triangular arrays

It is natural to wonder if central tendency will continue to hold for a (suitably normalised) sum of random variables with varying distributions. A small variant of the setting allows us to tackle situations where the individual variables have distributions that depend both on their position k in the sum and on the asymptotic parameter n itself. Subscripts are overloaded and we keep the dependence on n in the individual variables in view in a superscript notation.

Suppose m_n is a sequence of positive integers. For each n , suppose $X_1^{(n)}, \dots, X_{m_n}^{(n)}$ is a doubly indexed sequence of independent random variables on some probability space where, for each k , $X_k^{(n)}$ has distribution $F_k^{(n)}$ with zero mean $E(X_k^{(n)}) = 0$ and finite variance $\text{Var}(X_k^{(n)}) = (\sigma_k^{(n)})^2$ that may depend on n as well as on k . The doubly indexed sequence $\{X_k^{(n)}, 1 \leq k \leq m_n, n \geq 1\}$ is called a *triangular array*. The reason behind the name becomes apparent if we consider the case $m_n = n$ and arrange the elements row-wise in the form shown alongside. If, for each k , the variables $X_k^{(k)}, X_k^{(k+1)}, X_k^{(k+2)}, \dots$ in the k th column all have the same distribution then the situation reduces to that of a sequence of independent random variables $X_1, X_2, \dots, X_k, \dots$ with varying distributions $F_1, F_2, \dots, F_k, \dots$, respectively. In a generic triangular array we will only require that the terms in any given row are independent; the number m_n of terms in a row is not necessarily linear in n and across rows the variables may even be defined on different probability spaces.

We consider the row sums $S_n = X_1^{(n)} + \dots + X_{m_n}^{(n)}$. For each n , let $s_n^2 = (\sigma_1^{(n)})^2 + \dots + (\sigma_{m_n}^{(n)})^2$ be the variance of S_n . The random variable $S_n^* = S_n / s_n$ is then properly centred and normalised to zero mean and unit variance and it is natural to wonder if a limit law will hold for S_n^* . On consideration, it becomes apparent that some regularity condition will have to be placed on the marginal distributions $F_k^{(n)}$. The following example illustrates the kind of situation that may be troublesome.

$$\begin{aligned} & X_1^{(1)} \\ & X_1^{(2)}, X_2^{(2)} \\ & X_1^{(3)}, X_2^{(3)}, X_3^{(3)} \\ & X_1^{(4)}, X_2^{(4)}, X_3^{(4)}, X_4^{(4)} \\ & \cdots \cdots \cdots \end{aligned}$$

¹¹E. Bolthausen, “An estimate of the remainder in a combinatorial central limit theorem”, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 66, pp. 379–386, 1984.

EXAMPLE 1) Expansive uniform distributions. Suppose for $1 \leq k \leq n$ that $X_k^{(n)}$ has the uniform density $u_{-2^k, 2^k}(\cdot)$, that is to say, it is uniformly distributed in the interval $(-2^k, 2^k)$. Then its mean is zero and its variance is given by $\sigma_k^2 = 4^k/3$. The difficulty here appears to be that the variances are growing very quickly and, in particular, the variance of the final variable in the sum $S_n = X_1^{(n)} + \dots + X_n^{(n)}$ dominates. Indeed, $s_n^2 = \frac{1}{3} \sum_{k=1}^n 4^k = \frac{4}{9}(4^n - 1)$ whence $\sigma_n^2/s_n^2 = \frac{3}{4}(1 - \frac{1}{4^n}) \rightarrow \frac{3}{4}$ and $X_n^{(n)}$ by itself contributes about three-quarters of the variance of the sum. It appears difficult to find a common scale that would work in such a situation. To control $X_n^{(n)}$ we need to scale the sum by a factor of at least 2^n , the growth rate of the standard deviation σ_n , but such a scale would essentially squash most of the variables in the sum to a negligible contribution. If, on the other hand, we scale the variables by a factor which grows strictly slower than 2^n then $X_n^{(n)}$ “escapes” and its growth is uncontrolled. ►

In view of the previous example, if a limit theorem is to hold for the normalised row sum S_n^* of a triangular array it appears that we will need conditions to ensure that one variable (or a small group) do not dominate the sum. Lindeberg discovered the precise condition that is required.

$$\text{Lindeberg condition: } \frac{1}{s_n^2} \sum_{k=1}^{m_n} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x) \rightarrow 0 \quad (\text{for every } \epsilon > 0).$$

CENTRAL LIMIT THEOREM FOR TRIANGULAR ARRAYS *With notation as above, suppose the Lindeberg condition is satisfied. Then the normalised row sum $S_n^* = (X_1^{(n)} + \dots + X_{m_n}^{(n)})/s_n$ converges in distribution to the standard normal Φ .*

The Lindeberg condition looks to be excessively technical on the face of it—as if it were crafted for the sole reason of making things work out. In fact, as W. Feller discovered, it is also necessary. Before proceeding to the proof of the theorem a little further examination may help demystify things to some extent. For the individual variances, we have

$$\begin{aligned} (\sigma_k^{(n)})^2 &= \int x^2 dF_k^{(n)}(x) = \int_{|x| < \epsilon s_n} x^2 dF_k^{(n)}(x) + \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x) \\ &\leq \epsilon^2 s_n^2 + \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x). \end{aligned}$$

It follows that $(\sigma_k^{(n)})^2/s_n^2 \leq \epsilon^2 + \frac{1}{s_n^2} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x)$ and we hence obtain that

$$\max_{1 \leq k \leq m_n} \frac{(\sigma_k^{(n)})^2}{s_n^2} \leq \epsilon^2 + \frac{1}{s_n^2} \sum_{k=1}^{m_n} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x).$$

If the Lindeberg condition holds then the second term on the right tends to zero so that it is less than, say, ϵ^2 , eventually, and so $\max_{k \leq m_n} \sigma_k^{(n)} / s_n \leq \sqrt{2} \epsilon$ for all n sufficiently large. As ϵ may be chosen arbitrarily small, it follows that *if the Lindeberg condition holds then $\sigma_k^{(n)} / s_n \rightarrow 0$ uniformly for each $k \leq m_n$* . Thus, the effect of the Lindeberg condition is to ensure that any given variance is small compared to the sum variance. Roughly speaking, the normalised sum $S_n^* = \frac{1}{s_n} X_1^{(n)} + \dots + \frac{1}{s_n} X_{m_n}^{(n)}$ spreads matters out in a diffused fashion across the variables in the sum; no one variable has a decisive impact.

EXAMPLES: 2) *Variables of the same type.* Let F be any distribution with zero mean and unit variance. Suppose $\{X_k, k \geq 1\}$ is a sequence of independent random variables with X_k possessed of distribution $F_k(x) = F(x/\sigma_k)$ for each k . Then F_k has zero mean and variance σ_k^2 . With $m_n = n$ and $F_k^{(n)} = F_k$, if the Lindeberg condition is satisfied then $\sigma_k / s_n \rightarrow 0$ for $1 \leq k \leq n$. Conversely, suppose that $\max_{k \leq n} \sigma_k / s_n \rightarrow 0$. Then the change of variable $x \leftarrow y/\sigma_k$ shows that

$$\int_{|y| \geq \epsilon s_n} y^2 dF_k(y) = \sigma_k^2 \int_{|x| \geq \epsilon s_n / \sigma_k} x^2 dF(x).$$

For any $\epsilon > 0$ the integral on the right will eventually be less than ϵ because $\min_{k \leq n} s_n / \sigma_k \rightarrow \infty$ and the integral $\int x^2 dF(x)$ is convergent. It follows that

$$\frac{1}{s_n^2} \sum_{k=1}^n \int_{|x| \geq \epsilon s_n} x^2 dF_k(x) < \frac{\epsilon}{s_n^2} \sum_{k=1}^n \sigma_k^2 = \epsilon,$$

eventually. As $\epsilon > 0$ may be chosen arbitrarily small this implies that the Lindeberg condition is satisfied. Thus, *if the variables X_k are of the same type then the Lindeberg condition is satisfied if, and only if, $\max_{k \leq n} \sigma_k^2 / s_n^2 \rightarrow 0$* .

3) *Return to expansive uniform distributions.* The uniform distributions of Example 1 satisfy $\sigma_n^2 / s_n^2 \rightarrow 3/4$ so that the Lindeberg condition is violated. It is still possible to control the situation for expansive variances, however, if the variances are a little less explosive.

Suppose $X_k^{(n)}$ has the uniform distribution in the interval $(-k, k)$. Then F_k has mean zero and variance $k^2/3$ so that $s_n^2 = \frac{1}{3} \sum_{k=1}^n k^2 = n(n+1)(2n+1)/18 > n^3/9$. It follows that $\max_{k \leq n} \sigma_k^2 / s_n^2 = \sigma_n^2 / s_n^2 < 3/n \rightarrow 0$ and as the distributions F_k are of the same type the Lindeberg condition is satisfied.

4) *Bounded variables.* If the variables $X_k^{(n)}$ are uniformly bounded (in other words, there is some $A < \infty$ such that $\sup_{k,n} |X_k^{(n)}| \leq A$), and $s_n \rightarrow \infty$, then the Lindeberg condition is automatically satisfied. ►

Our notation now has to expand to accommodate the different marginal distributions. Let $F_{n,k}$ be the distribution of $X_k^{(n)} / s_n$. Then $F_{n,k}(y) = F_k^{(n)}(s_n y)$

has zero mean and variance $(\sigma_k^{(n)})^2/s_n^2$ tending to zero. We write $\mathfrak{F}_{n,k}$ for the operator associated with $F_{n,k}$. The key is the basic lemma of Section 1 of this chapter; we will need to do little more than replace \mathfrak{F}_n and the associated variances $1/n$ in the lemma by $\mathfrak{F}_{n,k}$ and $(\sigma_k^{(n)})^2/s_n^2$ to reflect the altered circumstances.

THE BASIC LEMMA II *Suppose u is a continuous function with three bounded derivatives. Then*

$$\sum_{k=1}^{m_n} \left\| \mathfrak{F}_{n,k} u - u - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u'' \right\| \rightarrow 0.$$

PROOF: With $q_t(y)$ as defined in (1.1), the relation (1.2) must now amend to

$$\mathfrak{F}_{n,k} u(t) - u(t) - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u''(t) = \int_{-\infty}^{\infty} q_t(y) dF_{n,k}(y).$$

We now proceed to use the Taylor expansion of $u(t-y)$ around t to bound $q_t(y)$ in precisely the same way as before and arrive at the following counterpart of (1.3):

$$\left| \mathfrak{F}_{n,k} u(t) - u(t) - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u''(t) \right| \leq M\epsilon \frac{(\sigma_k^{(n)})^2}{s_n^2} + \frac{M}{s_n^2} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x).$$

Summing over k we hence obtain

$$\sum_{k=1}^{m_n} \left| \mathfrak{F}_{n,k} u(t) - u(t) - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u''(t) \right| \leq M\epsilon + \frac{M}{s_n^2} \sum_{k=1}^{m_n} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x).$$

The second term on the right tends to zero independently of the choice of t by the Lindeberg condition. As $\epsilon > 0$ may be chosen arbitrarily small, the claimed result follows. ▶

The steps now mirror the proof for the case of identical distributions. Suppose $Z_1^{(n)}, \dots, Z_{m_n}^{(n)}$ are independent with $Z_k^{(n)}$ possessed of the normal distribution Φ_k with mean zero and variance $(\sigma_k^{(n)})^2$. Write $T_n^* = (Z_1^{(n)} + \dots + Z_{m_n}^{(n)})/s_n$. Let $\Phi_{n,k}$ denote the distribution of $Z_k^{(n)}/s_n$, whence $\Phi_{n,k}(y) = \Phi_k(s_n y)$, with $\mathfrak{N}_{n,k}$ the associated operator. Then T_n^* has distribution given by the n -fold convolution $\Phi_{n,1} * \dots * \Phi_{n,n}$ with corresponding product operator $\mathfrak{N}_{n,1} * \dots * \mathfrak{N}_{n,n}$. As the normal distribution is closed under convolution it is clear that T_n^* is, in fact, a standard normal random variable whence $\Phi = \Phi_{n,1} * \dots * \Phi_{n,n}$, or equivalently, $\mathfrak{N} = \mathfrak{N}_{n,1} * \dots * \mathfrak{N}_{n,n}$. Of course, the distribution of S_n^* is given by the n -fold convolution $F_{n,1} * \dots * F_{n,n}$ with corresponding

product operator $\mathfrak{F}_{n,1} \cdots \mathfrak{F}_{n,n}$. The idea, as before, is to compare the distribution of S_n^* with that of T_n^* , or equivalently, to compare the operators $\mathfrak{F}_{n,1} \cdots \mathfrak{F}_{n,n}$ and $\mathfrak{N}_{n,1} \cdots \mathfrak{N}_{n,n}$.

PROOF OF THE THEOREM: Let u be any function with three bounded derivatives. Again, as a consequence of the reductionist theorem of Section XIX.5, we have

$$\|\mathfrak{F}_{n,1} \cdots \mathfrak{F}_{n,n} u - \mathfrak{N}_{n,1} \cdots \mathfrak{N}_{n,n} u\| \leq \sum_{k=1}^{m_n} \|\mathfrak{F}_{n,k} u - \mathfrak{N}_{n,k} u\|.$$

Adding and subtracting the term $u + \frac{(\sigma_k^{(n)})^2}{2s_n^2} u''$ inside the norm on the right, the triangle inequality shows that the right-hand side is bounded above by

$$\sum_{k=1}^{m_n} \left\| \mathfrak{F}_{n,k} u - u - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u'' \right\| + \sum_{k=1}^{m_n} \left\| \mathfrak{N}_{n,k} u - u - \frac{(\sigma_k^{(n)})^2}{2s_n^2} u'' \right\|,$$

and each of the terms above tends to zero in view of the just-proved lemma. It follows that $\|\mathfrak{F}_{1,n} \cdots \mathfrak{F}_{n,n} u - \mathfrak{N}_n u\| \rightarrow 0$ for all functions u with three bounded derivatives, hence for all smooth functions, which is the same as saying that $F_{1,n} \star \cdots \star F_{n,n} \xrightarrow{v} \Phi$. ▶

In practice the Lindeberg condition can be difficult to verify. In 1900, A. M. Liapounov proved the central limit theorem under the following slightly more demanding condition.

$$\text{Liapounov condition: } \frac{1}{s_n^r} \sum_{k=1}^{m_n} E(|X_k^{(n)}|^r) \rightarrow 0 \quad (\text{for some } r > 2).$$

Problem 8 provides an example where the Lindeberg condition holds but the Liapounov condition does not.

To see that Liapounov's condition is in fact more restrictive than Lindeberg's condition, we observe that, if $r > 2$,

$$E(|X_k^{(n)}|^r) \geq \int_{|x| \geq \epsilon s_n} |x|^r dF_k^{(n)}(x) \geq \epsilon^{r-2} s_n^{r-2} \int_{|x| \geq \epsilon s_n} |x|^2 dF_k^{(n)}(x)$$

as $|x|^r = |x|^2 \cdot |x|^{r-2} \geq |x|^2 \epsilon^{r-2} s_n^{r-2}$ if $|x| \geq \epsilon s_n$. It follows that

$$\frac{1}{s_n^2} \sum_{k=1}^{m_n} \int_{|x| \geq \epsilon s_n} x^2 dF_k^{(n)}(x) \leq \frac{1}{\epsilon^{r-2} s_n^r} \sum_{k=1}^{m_n} E(|X_k^{(n)}|^r)$$

and the Lindeberg condition will be satisfied if the Liapounov condition holds. We thus obtain the following

COROLLARY If the Liapounov condition is satisfied then S_n^* converges in distribution to the standard normal Φ .

The utility of this variant is computational; the Liapounov condition is often easier to verify than the Lindeberg condition as moments, typically, are easier to estimate than tails. The reader will find an example in the return to a classical theme in the following section.

6 The coupon collector

When variables are unbounded, calculation of probability tails can get messy. In these situations moments are frequently easier to compute. A classical example serves to reinforce this point.

Our indefatigable coupon collector wishes to collect a fixed fraction $0 < \rho < 1$ of n novelty items. Let $m_n = \lfloor \rho n \rfloor$. Then m_n differs from ρn by no more than one and $m_n/n \rightarrow \rho$. Suppose that she makes S_n purchases before collecting m_n distinct coupons. What can we say about the asymptotic distribution of S_n ?

Assuming that the coupon varieties are uniformly distributed in a very large pool of items, the situation corresponds to that of random sampling from n coupons with replacement. Let $X_k^{(n)}$ be the waiting time (the number of purchases) between the times when the $(k-1)$ th and k th distinct new coupons are first obtained. If $k-1$ distinct coupons have already been obtained then a new coupon is obtained on a given purchase with probability $p = (n-k+1)/n$. It follows that $X_k^{(n)}$ has a geometric distribution with mean waiting time $(1-p)/p = (k-1)/(n-k+1)$ and variance $(1-p)/p^2 = n(k-1)/(n-k+1)^2$. Furthermore, as sampling is with replacement, the variables $X_1^{(n)}, \dots, X_{m_n}^{(n)}$ are independent and account for the “wasted” purchases in between the acquisition of previously unseen coupons. It follows that $S_n = m_n + X_1^{(n)} + \dots + X_{m_n}^{(n)}$ is a row sum of a triangular array. By additivity of expectation,

$$E(S_n) = m_n + \sum_{k=1}^{m_n} E(X_k^{(n)}) = m_n + \sum_{k=1}^{m_n} \frac{k-1}{n-k+1} = \sum_{k=1}^{m_n} \frac{1}{1 - \frac{k-1}{n}}.$$

Evaluating the sum on the right by integrals, we obtain

$$\int_0^{m_n} \frac{dt}{1 - \frac{t}{n}} < \sum_{k=1}^{m_n} \frac{1}{1 - \frac{k-1}{n}} < 1 + \int_1^{m_n+1} \frac{dt}{1 - \frac{t}{n}}.$$

Changing the variable of integration to $x = t/n$, an easy evaluation of integrals shows that each of the integrals on the left and right differs from $n \int_0^\rho \frac{dx}{1-x}$ only

by a bounded amount and, consequently,

$$E(S_n) = n \int_0^\rho \frac{dx}{1-x} + \xi_n = -n \log(1-\rho) + \xi_n$$

where $\xi_n = O(1)$ is a sequence of bounded values. Likewise,

$$s_n^2 = \text{Var}(S_n) = \sum_{k=1}^{m_n} \text{Var}(X_k^{(n)}) = \sum_{k=1}^{m_n} \frac{n(k-1)}{(n-k+1)^2} = n \sum_{j=0}^{m_n-1} \frac{j}{(n-j)^2},$$

by additivity of variance. The function $f(x) = x/(n-x)^2$ increases monotonically in the range $0 < x < n$ as is easily verified by taking derivatives and observing that $f'(x) = (n+x)/(n-x)^3$ is strictly positive for $0 < x < n$. We may hence bound the sum on the right by integrals, obtaining

$$\int_0^{m_n-1} \frac{t}{(n-t)^2} dt < \sum_{j=0}^{m_n-1} \frac{j}{(n-j)^2} < \int_1^{m_n} \frac{t}{(n-t)^2} dt.$$

Again, changing the variable of integration to $x = t/n$, evaluating the easy integrals by parts, and collecting subdominant terms, shows readily that

$$s_n^2 = n \int_0^\rho \frac{x}{(1-x)^2} dx + \zeta_n = nh(\rho) + \zeta_n$$

where $\zeta_n = O(1)$ is another sequence of bounded values and the function h is defined by

$$h(\rho) = \frac{1}{1-\rho} [\rho + (1-\rho) \log(1-\rho)] \quad (0 \leq \rho < 1).$$

(It is easy to see that $h(\rho)$ is strictly positive in the range $0 < \rho < 1$ as $h(0) = 0$ and $[(1-\rho)h(\rho)]' = -\log(1-\rho) > 0$ for $0 < \rho < 1$.) Thus, $s_n \rightarrow \infty$ and a central limit theorem now beckons.

We begin with an easy algebraic manoeuvre. In view of the trite inequality $2|ab| \leq a^2 + b^2$, we have $(a \pm b)^2 \leq 2(a^2 + b^2)$. A repeated application yields $(a-b)^4 = (a-b)^2(a-b)^2 \leq 4(a^2 + b^2)^2 \leq 8(a^4 + b^4)$. (The reader may recognise an elementary form of Hölder's inequality.) Now suppose X is a geometric random variable with mean $q/p = (1-p)/p$ for some $0 < p \leq 1$. Then $E((X - q/p)^4) \leq 8E(X^4) + 8q^4/p^4 \leq 8E(X^4) + 8/p^4$ and it suffices to estimate the fourth moment of a geometric variable.

As X^4 is an integrable, positive random variable, by virtue of Theorem XIII.5.4 (see also Problem XIII.3), we have

$$\begin{aligned} E(X^4) &= \int_0^\infty P\{X^4 \geq t\} dt = \int_0^\infty P\{X \geq t^{1/4}\} dt = \frac{4}{p^4} \int_0^\infty P\{X \geq u/p\} u^3 du \\ &= \frac{4}{p^4} \int_0^\infty (1-p)^{\lceil u/p \rceil} u^3 du \leq \frac{4}{p^4} \int_0^\infty (1-p)^{u/p} u^3 du. \end{aligned}$$

Normal Approximation

The reader may recall the elementary inequality $(1 - p)^\nu \leq e^{-p\nu}$, valid for all $0 < p \leq 1$ and $\nu \geq 0$. (If she doesn't, she can readily verify it by taking logarithms and recalling from the Taylor series for the logarithm that $\log(1 - p) = -p - p^2/2 - p^3/3 - \dots \leq -p$.) It follows that

$$\mathbb{E}(X^4) \leq \frac{4}{p^4} \int_0^\infty e^{-u} u^3 du = \frac{24}{p^4}$$

by an easy integration by parts. Putting the pieces together, we have the estimate $\mathbb{E}((X - q/p)^4) \leq 32/p^4$ for the centred fourth moment of a geometric variable.

As $S_n - \mathbb{E}(S_n) = \sum_{k=1}^{m_n} (X_k^{(n)} - \frac{k-1}{n-k+1})$, it will suffice to verify the Liapounov condition for the centred geometric variables $X_k^{(n)} - \frac{k-1}{n-k+1}$. We have

$$\begin{aligned} \sum_{k=1}^{m_n} \mathbb{E} \left[\left(X_k^{(m)} - \frac{k-1}{n-k+1} \right)^4 \right] &\leq 32 \sum_{k=1}^{m_n} \left(\frac{n}{n-k+1} \right)^4 \\ &= 32 \sum_{j=0}^{m_n-1} \frac{1}{\left(1 - \frac{j}{n}\right)^4} = 32n \int_0^\rho \frac{dx}{(1-x)^4} + \eta_n \end{aligned}$$

where we again estimate the sum in the penultimate step by an integral, the error terms $\eta_n = \mathcal{O}(1)$ forming a bounded sequence. The integral is easy enough to evaluate but its precise value makes little matter: we have

$$\frac{1}{s_n^4} \sum_{k=1}^{m_n} \mathbb{E} \left[\left(X_k^{(m)} - \frac{k-1}{n-k+1} \right)^4 \right] \leq \frac{\frac{32n}{3} \left(\frac{1}{(1-\rho)^3} - 1 \right) + \eta_n}{(nh(\rho) + \zeta_n)^2} \leq \frac{C}{n}$$

for an absolute positive constant $C = C(\rho)$ determined by ρ , and the Liapounov condition holds. (The reader may feel that the choice of power 4 is vaguely familiar. If she revisits the proof of Borel's normal law in Section V.7 she will see why.)

As usual, let $S_n^* = (S_n - \mathbb{E}(S_n))/s_n$ be the normalised row sum. Then S_n^* converges in distribution to a standard normal random variable Z . Now,

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{s_n} = \frac{S_n + n \log(1 - \rho) - \xi_n}{\sqrt{nh(\rho)} \left(1 + \frac{\zeta_n}{nh(\rho)}\right)^{1/2}}.$$

Reusing notation and setting $a_n = \left(1 + \frac{\zeta_n}{nh(\rho)}\right)^{1/2}$ and $b_n = \xi_n/\sqrt{nh(\rho)}$, we obtain

$$\frac{S_n + n \log(1 - \rho)}{\sqrt{nh(\rho)}} = a_n S_n^* + b_n.$$

As $n \rightarrow \infty$, it is clear that $a_n \rightarrow 1$ and $b_n \rightarrow 0$, leading to a neat asymptotic description.

THEOREM Let S_n be the number of purchases made to acquire $\lfloor \rho n \rfloor$ distinct coupons where $0 < \rho < 1$. Then $(S_n + n \log(1 - \rho)) / \sqrt{n h(\rho)}$ converges in distribution to a standard normal random variable Z .

Roughly speaking, $S_n = -n \log(1 - \rho) + O(\sqrt{n})$, and the “inefficiency” $S_n/\rho n$ representing the number of excess purchases per new coupon is approximately $-\log(1 - \rho)/\rho$.

7 On the number of cycles

A pretty illustration of the ubiquity of the triangular array setting can be seen in the following combinatorial example.

Suppose $k \mapsto \pi(k)$ is a permutation of the numbers from 1 to n . Beginning with $\pi^1(k) = k$, it will be useful to compact notation and set $\pi^2(k) = \pi(\pi(k))$, $\pi^3(k) = \pi(\pi(\pi(k)))$, and so on, with π^j representing the j th iterate of π . It will also be convenient to write $\pi^0(k) = k$ for the identity map. If, starting from 1, we follow the permutation iterates $1 = \pi^0(1) \mapsto \pi^1(1) \mapsto \pi^2(1) \mapsto \pi^3(1) \mapsto \dots$, then, sooner or later, we will complete a cycle $\pi^j(1) = 1$ for some j and further iterates of the subsequence $(\pi^0(1), \pi^1(1), \dots, \pi^{j-1}(1))$ will only cycle through these j numbers in sequence. If we now begin anew with the smallest integer not yet encountered we will build up a new cycle and, proceeding hence, by sequentially listing the elements of each cycle and beginning anew each time a cycle is completed with the smallest integer not yet encountered, we can systematically group the numbers from 1 to n into a collection of cycles induced by π . The permutation of the numbers from 1 to 7 shown below illustrates the process.

$$\begin{array}{c|ccccccc} k & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline \pi(k) & 4 & 5 & 7 & 1 & 3 & 6 & 2 \end{array} \quad 1 \mapsto 4 \mapsto 1, \quad 2 \mapsto 5 \mapsto 3 \mapsto 7 \mapsto 2, \quad 6 \mapsto 6 \\ (1 \ 4) \quad (2 \ 5 \ 3 \ 7) \quad (6).$$

As we see, there are three cycles corresponding to the given permutation π (parentheses added for visual effect only). Each grouping lists the elements of a cycle in sequence, from the first element to the last, before the cycle starts anew, leading to the cyclical rearrangement $((1, 4), (2, 5, 3, 7), (6))$ with opening parentheses indicating the start of a new cycle, and closing parentheses showing where cycles are terminated.

Suppose now that $k \mapsto \Pi(k)$ is a random permutation of the integers 1 through n . What can we say about the number of cycles engendered by it? Consider the sequence of indicators $X_1^{(n)}, \dots, X_n^{(n)}$ where, for each k , $X_k^{(n)}$ is the indicator for whether location k in the cyclical arrangement completes a cycle. Then $S_n = X_1^{(n)} + \dots + X_n^{(n)}$ is the number of cycles engendered by Π . In the above example, $X_2^{(7)} = X_6^{(7)} = X_7^{(7)} = 1$ and $X_1^{(7)} = X_3^{(7)} = X_4^{(7)} = X_5^{(7)} = 0$, so that the number of cycles is $S_7 = 3$.

We begin with a consideration of the distribution of the variables $X_1^{(n)}, \dots, X_n^{(n)}$. It is clear that $X_1^{(n)} = 1$ if, and only if, $\Pi(1) = 1$, an event of probability $1/n$. Let us consider now the situation at location k in the cyclical arrangement. Preceding k there must be a *largest* integer $j < k$ at which location the previous cycle was completed. (We set $j = 0$ if no prior cycle has been completed before k .) Suppose α_0 is the smallest unused integer through the first j locations in the cycle-based rearrangement. Then a new cycle begins at location $j + 1$ with α_0 and, by definition of j , is not terminated through location $k - 1$. In other words, if we write $\alpha_j = \Pi^j(\alpha_0)$, then the sequence of iterates $\alpha_0, \alpha_1, \dots, \alpha_{k-j-2}$ from the $(j + 1)$ th location to the $(k - 1)$ th location has no repeated element and $\Pi(\alpha_{k-j-2}) = \alpha_{k-j-1} \neq \alpha_0$. After the specification of the first $k - 1$ elements in the cyclical arrangement, there will be a total of $n - k + 1$ unused integers of which one is α_{k-j-1} (as the cycle beginning with α_0 has not been terminated). Conditioned on the sequence through the $(k - 1)$ th location, the cycle will terminate at location k hence if, and only if, $\Pi(\alpha_{k-j-1}) = \Pi^{k-j}(\alpha_0) = \alpha_0$, an event of probability $1/(n - k + 1)$. As this probability does not depend upon the location j where the previous cycle was completed, or the particular sequence through the first $k - 1$ elements of the cyclical rearrangement, it follows, in particular, that the variables $X_1^{(n)}, \dots, X_n^{(n)}$ are independent, and, by taking expectation with respect to the first $k - 1$ elements to remove the conditioning, we obtain $P\{X_k^{(n)} = 1\} = 1/(n - k + 1)$. Thus, $X_1^{(n)}, \dots, X_n^{(n)}$ forms a sequence of Bernoulli trials with $X_k^{(n)}$ representing the outcome of a coin toss with success probability $1/(n - k + 1)$.

It is always wise to look backwards to verify our calculations. If $k = 1$ then $X_1^{(n)}$ is a Bernoulli trial with success probability $1/n$, as observed earlier. And, as the final element of the cyclical arrangement *must* complete a cycle (else there will be an unconsummated element left over), at the other end, by setting $k = n$, we see indeed that $X_n^{(n)}$ is a Bernoulli trial with certain success, $X_n^{(n)} = 1$.

It follows hence that $S_n = X_1^{(n)} + \dots + X_n^{(n)}$ represents a row sum of a triangular array. By additivity, its expectation is given by $E(S_n) = 1 + \frac{1}{2} + \dots + \frac{1}{n}$. As the function $1/x$ decreases monotonically, we may estimate the sum on the right by integrals,

$$\int_1^{n+1} \frac{dx}{x} < \sum_{j=1}^n \frac{1}{j} < 1 + \int_1^n \frac{dx}{x},$$

and evaluating the elementary integrals on either side shows that $E(S_n) = \log n + \xi_n$ where the sequence of values ξ_n is bounded in absolute value. Likewise, as the variables $X_1^{(n)}, \dots, X_n^{(n)}$ are independent, additivity of variance shows that

$$s_n^2 = \text{Var}(S_n) = \sum_{k=1}^n \frac{1}{n-k+1} \left(1 - \frac{1}{n-k+1}\right) = \sum_{j=1}^n \frac{1}{j} - \sum_{j=1}^n \frac{1}{j^2}.$$

As the function $1/x^2$ is monotonically decreasing, we may again bound the second sum on the right by integrals,

$$\int_1^{n+1} \frac{dx}{x^2} < \sum_{j=1}^n \frac{1}{j^2} < 1 + \int_1^n \frac{dx}{x^2},$$

and, evaluating the easy integrals shows that $s_n^2 = \log n + \zeta_n$ where the sequence of values ζ_n is again bounded in absolute value. As $s_n \rightarrow \infty$ (if slowly) and the centred Bernoulli variables $X_1^{(n)} - 1/n, \dots, X_k^{(n)} - 1/(n-k+1), \dots, X_n^{(n)} - 1$ are uniformly bounded, Lindeberg's condition holds tritely and the central limit theorem is in force. The centred and normalised row sums

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - \log n - \xi_n}{\sqrt{\log n} \left(1 + \frac{\zeta_n}{\log n}\right)^{1/2}}$$

hence converge in distribution to the standard normal Φ . Setting $a_n = (1 + \zeta_n/\log n)^{1/2}$ and $b_n = \xi_n/\sqrt{\log n}$, we have

$$\frac{S_n - \log n}{\sqrt{\log n}} = a_n S_n^* + b_n,$$

and, as in the previous section, we have $a_n \rightarrow 1$ and $b_n \rightarrow 0$, an elegant result beckons (see Problem XIX.1).

THEOREM *Let S_n be the number of cycles engendered by a random permutation of the integers from 1 to n , Z any standard normal. Then $(S_n - \log n)/\sqrt{\log n} \xrightarrow{d} Z$.*

Roughly speaking, the number of cycles generated by a random permutation is $\log n + O(\sqrt{\log n})$. This is Goncharov's theorem.¹²

8 Many dimensions

Another direction where the basic approach generalises effortlessly is when the variables take values in more than one dimension. The notation gets inescapably more cumbersome but vector notation helps keep the mess under control. For the rest of this section, we use lowercase bold face letters such as \mathbf{x} , \mathbf{y} , \mathbf{s} , and \mathbf{t} to represent points in a v -dimensional real Euclidean space with subscripts serving to identify their components in the standard basis. Thus, $\mathbf{x} = (x_1, \dots, x_v)$ is a point in \mathbb{R}^v with components x_1, \dots, x_v in each of the v orthogonal directions. In this representation, it will be convenient to think of

¹²V. Goncharov, "Du domaine d'analyse combinatoire", *Bulletin de l'Académie Sciences URSS, Sér. Math.*, vol. 8, pp. 3–48, 1944.

\mathbf{x} as a *row* vector. In keeping with convention, random vectors in \mathbb{R}^v are then represented by uppercase bold face letters such as \mathbf{X} , \mathbf{Y} , and \mathbf{S} .

Let F be any distribution in v dimensions, $F(\mathbf{x}) = F(x_1, \dots, x_v)$. We will suppose that F is centred, that is to say, it has zero mean with $\int \mathbf{x} dF(\mathbf{x}) = 0$, and has a non-degenerate covariance matrix $\Sigma = [\sigma_{i,j}] = \int \mathbf{x}^\top \mathbf{x} dF(\mathbf{x})$. Of course, all integrals are over the space \mathbb{R}^v and expectations of vectors and matrices are to be interpreted componentwise. For the purposes of this section we introduce the nonce notation Φ_Σ for the normal distribution with mean zero and covariance matrix Σ in v dimensions. As we've seen in Section X.4, Φ_Σ is the distribution corresponding to the centred normal density

$$\phi_\Sigma(\mathbf{x}) = \frac{1}{(2\pi)^{v/2}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1} \mathbf{x}\right)$$

in v dimensions. We will not need the explicit form of the density in what follows; as in the one-dimensional case, the only key fact about the normal that we need appeal to is that the sum of independent normals in any number of dimensions is normal with mean vector and covariance matrix equal to the sum of the constituent means and covariance matrices, respectively.

CENTRAL LIMIT THEOREM FOR MANY DIMENSIONS Suppose $\mathbf{X}_1, \mathbf{X}_2, \dots$ are independent random variables in v dimensions, each \mathbf{X}_k drawn from the common distribution F with zero mean and non-degenerate covariance matrix Σ . Let $\mathbf{S}_n^* = (\mathbf{X}_1 + \dots + \mathbf{X}_n)/\sqrt{n}$. Then \mathbf{S}_n^* converges in distribution to the normal distribution Φ_Σ .

Note that it is not necessary in the theorem that the components of the random vector $\mathbf{X}_k = (X_{k,1}, \dots, X_{k,v})$ be independent; indeed, the v components $X_{k,1}, \dots, X_{k,v}$ can have any dependency structure whatsoever subject only to the constraint that the covariance matrix Σ be non-degenerate.

Vector notation will help compact expressions. If $u(\mathbf{t}) = u(t_1, \dots, t_v)$ is a thrice differentiable function, we write $u_i(\mathbf{t})$, $u_{i,j}(\mathbf{t})$, and $u_{i,j,k}(\mathbf{t})$ for the partial derivatives of u with respect to the components t_i, t_j, t_k indicated in the subscripts, $\mathcal{D}_1 u$ for the row vector of componentwise derivatives (u_1, \dots, u_v) , and $\mathcal{D}_2 u$ for the symmetric $v \times v$ matrix of second derivatives $[u_{i,j}]$. Also, if \mathbf{X} is equipped with distribution F , we write $\mathbb{S}u$ for the expectation of $\mathbf{X}(\mathcal{D}_2 u)\mathbf{X}^\top$ with respect to F , that is to say,

$$\mathbb{S}u(\mathbf{t}) = \int \mathbf{x}(\mathcal{D}_2 u(\mathbf{t})) \mathbf{x}^\top dF(\mathbf{x}) = \sum_{i,j} \sigma_{i,j} u_{i,j}(\mathbf{t})$$

where $\Sigma = [\sigma_{i,j}]$ is the covariance matrix of \mathbf{X} .

The supporting notation that we will need mirrors the earlier analysis. If \mathbf{X} has distribution F , we again write F_n for the distribution of \mathbf{X}/\sqrt{n} and let \mathfrak{F}_n be the associated convolutional operator. We then have the following analogue of the basic lemma in v dimensions.

THE BASIC LEMMA III *Suppose u is a continuous function of v variables and suppose that all partial derivatives of u through order three exist and are bounded. Then*

$$n \|\mathfrak{F}_n u - u - \frac{1}{2n} \mathcal{S}u\| \rightarrow 0.$$

PROOF: We redefine the function of (1.1) by

$$q_t(\mathbf{y}) = u(t - \mathbf{y}) - u(t) + (\mathcal{D}_1 u(t)) \mathbf{y}^\top - \frac{1}{2} \mathbf{y} (\mathcal{D}_2 u(t)) \mathbf{y}^\top$$

to reflect the changes appropriate to a move to v dimensions. Via the change of variable $\mathbf{x} = \sqrt{n} \mathbf{y}$, additivity of expectation now shows that

$$\int_{\mathbb{R}^v} q_t(\mathbf{y}) dF_n(\mathbf{y}) = \mathfrak{F}_n u(t) - u(t) - \frac{1}{2n} \mathcal{S}u(t). \quad (8.1)$$

We proceed to exploit the concentration of F_n near the origin as before.

As u has three bounded derivatives, we can find a positive quantity M which bounds from above the absolute value of each of the partial derivatives of u through order three. Fix any $\epsilon > 0$ and let $\mathbb{I}_\epsilon = (-\epsilon, \epsilon)$. Then $\mathbb{I}_\epsilon^v = \mathbb{I}_\epsilon \times \cdots \times \mathbb{I}_\epsilon$ represents the axis-parallel rectangle of side 2ϵ centred at the origin in \mathbb{R}^v and, in view of the concentration of F_n at the origin, $F_n(\mathbb{I}_\epsilon^v) > 1 - \epsilon$, eventually. We now partition the domain of integration \mathbb{R}^v in (8.1) into \mathbb{I}_ϵ^v and $(\mathbb{I}_\epsilon^v)^c$ and evaluate the contributions in turn.

For $\mathbf{y} \in \mathbb{I}_\epsilon^v$, truncating the multivariate Taylor expansion of $u(t - \mathbf{y})$ around t to three terms shows that $q_t(\mathbf{y}) = \frac{1}{6} \sum_{i,j,k} u_{i,j,k}(\xi) y_i y_j y_k$ for some $\xi = \xi(t, \mathbf{y})$. Simple bounds suffice here. As $|y_k| < \epsilon$ for each k , we have

$$|q_t(\mathbf{y})| \leq \frac{1}{6} M v \epsilon \sum_{i,j} |y_i y_j| \leq \frac{1}{12} M v \epsilon \sum_{i,j} (y_i^2 + y_j^2)$$

as $|ab| \leq \frac{1}{2}(a^2 + b^2)$ by the inequality of arithmetic and geometric means [or simply because $(a \pm b)^2 \geq 0$]. It follows that $|q_t(\mathbf{y})| \leq \frac{1}{6} M v \epsilon \|\mathbf{y}\|^2$ for all $\mathbf{y} \in \mathbb{I}_\epsilon^v$. (In standard notation, here $\|\mathbf{y}\|^2 = y_1^2 + \cdots + y_v^2$ denotes the square of the Euclidean vector norm; as always, the context tells us which norm is in operation.) Consequently,

$$\left| \int_{\mathbb{I}_\epsilon^v} q_t(\mathbf{y}) dF_n(\mathbf{y}) \right| \leq \frac{1}{6} M v \epsilon \int_{\mathbb{I}_\epsilon^v} \|\mathbf{y}\|^2 dF_n(\mathbf{y}) \leq \frac{1}{6} M v \epsilon \int_{\mathbb{R}^v} \|\mathbf{y}\|^2 dF_n(\mathbf{y}) = C_1 \frac{\epsilon}{n}$$

for an absolute positive constant $C_1 = \frac{1}{6} M v (\sigma_{1,1} + \cdots + \sigma_{v,v})$ independent of t and ϵ .

When $\mathbf{y} \notin \mathbb{I}_\epsilon^v$, a similar Taylor expansion of $u(t - \mathbf{y})$ through two terms shows that

$$q_t(\mathbf{y}) = \frac{1}{2} \mathbf{y} \mathcal{D}_2 u(\eta) \mathbf{y}^\top - \frac{1}{2} \mathbf{y} \mathcal{D}_2 u(t) \mathbf{y}^\top$$

for some $\eta = \eta(t, y)$. In view of the boundedness of the second derivatives, each term on the right is no more than $\frac{1}{2}M\|y\|^2$ in absolute value and it follows that $|q_t(y)| \leq M\|y\|^2$. Writing $\mathbb{I}_{\epsilon\sqrt{n}}^v$ for the axis-parallel box of side $2\epsilon\sqrt{n}$ in \mathbb{R}^v , via the usual change of variable $x = \sqrt{n}y$, we then obtain

$$\left| \int_{(\mathbb{I}_\epsilon^v)^c} q_t(y) dF_n(y) \right| \leq M \int_{y \notin \mathbb{I}_\epsilon^v} \|y\|^2 dF_n(y) = \frac{M}{n} \int_{x \notin \mathbb{I}_{\epsilon\sqrt{n}}^v} \|x\|^2 dF(x).$$

As $\int_{\mathbb{R}^v} \|x\|^2 dF(x)$ converges to $\sigma_{1,1} + \dots + \sigma_{v,v}$, the integral on the right decreases to zero as $n \rightarrow \infty$. In particular, for all sufficiently large n ,

$$\left| \int_{(\mathbb{I}_\epsilon^v)^c} q_t(y) dF_n(y) \right| \leq C_2 \frac{\epsilon}{n}$$

for an absolute positive constant C_2 independent of t and ϵ .

Pooling estimates we have $|\mathfrak{F}_n u(t) - u(t) - \frac{1}{2n} \mathcal{S}u(t)| \leq (C_1 + C_2) \frac{\epsilon}{n}$ for all sufficiently large n , the rate of convergence independent of t . It follows that $n(\mathfrak{F}_n u(t) - u(t) - \frac{1}{2n} \mathcal{S}u(t))$ converges uniformly to zero. ▶

The proof of the central limit theorem for identical distributions adapts directly to this setting in view of the above lemma and no further comment is needed. (If the reader is not convinced she should attempt to supply the details for herself in the style of the earlier proofs.)

As an alternative to operator methods, in this context, Fourier methods were proposed by A. Markov and fully exploited by P. Lévy—the reader has seen the idea in action in Section 2 of this chapter and earlier in Sections VI.4,5—and, especially before Lindeberg’s proof was fully understood and simplified, became the standard approach to the central limit theorem. Where applicable, Fourier methods will give sharper and deeper results than the operator methods we’ve used hitherto. The cost is in the increased sophistication of the tool. The interested reader may wish to consult the influential monograph of R. N. Bhattacharya and R. R. Rao which provides a panoramic survey of a variety of topics on the subject of normal approximation.¹³

Quo vadis? We could seek extensions of central tendency to handle large deviations in the style of Section VI.7. An example of the kind of extension that is feasible was provided by W. Richter¹⁴ using saddle point methods to evaluate a Fourier integral; Richter’s approach was extended by S. C. Fang and S. S. Venkatesh to handle large deviations in triangular arrays—and as a nice

¹³R. N. Bhattacharya and R. R. Rao, *Normal Approximation and Asymptotic Expansions*. Philadelphia, PA: SIAM, Classics in Applied Mathematics, 2010.

¹⁴W. Richter, “Multidimensional local limit theorems for large deviations”, *Theory of Probability and its Applications*, vol. 3, pp. 100–106, 1958.

bonus, the result put to work in a problem in binary integer programming.¹⁵ On another front we could seek refinements of the central limit theorem when the summands are dependent. Stein's subtle method of approximation which was put to such good effect in Poisson approximation in Chapter XVIII was originally developed for this purpose—and appears in a cameo in Sections 3 and 4.¹⁶ As in the case of Poisson approximation for dependent summands this is an area rich with possibility and has attracted much interest. I shall leave these tempting avenues to the reader to explore in the literature and proceed directly to applications.

9 Random walks, random flights

To illustrate the multivariate central limit theorem in action we consider two problems in random walks; the settings are familiar, the viewpoint new.

EXAMPLE 1) *Random walks in two and more dimensions.* Consider a random walk $S_n = X_1 + \dots + X_n$ in the Euclidean plane \mathbb{R}^2 . Each $X_k = (X_{k,1}, X_{k,2})$ represents a random, axis-parallel step in the plane; if $e_1 = (1, 0)$ and $e_2 = (0, 1)$ are the standard basis elements then X_k takes one of the four values $\pm e_1, \pm e_2$, each with equal probability $1/4$. It is clear then that the components $X_{k,1}$ and $X_{k,2}$ take values ± 1 , each with probability $1/4$, and value 0 with probability $1/2$. Thus, each of the components has mean zero and variance $1/2$. On the other hand, precisely one of $X_{k,1}$ and $X_{k,2}$ is non-zero so that $\text{Cov}(X_{k,1}, X_{k,2}) = 0$ or, in words, $X_{k,1}$ and $X_{k,2}$ are uncorrelated. It follows that $\Sigma = \text{Cov}(X_k) = \frac{1}{2}I_2$ where $I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ is the 2×2 identity matrix. Note that while the components $X_{k,1}$ and $X_{k,2}$ are uncorrelated, they are manifestly *dependent*. The multivariate central limit theorem now tells us that S_n/\sqrt{n} converges in distribution to the two-dimensional, centred normal distribution with covariance matrix $\frac{1}{2}I_2$. However, uncorrelated (joint) normals are independent. It follows that S_n^* has asymptotically independent components distributed as marginal normals with mean zero and variance $1/2$.

Everything carries over to a walk S_n in v dimensions. Let e_1, \dots, e_v represent the unit vectors in each of the v coordinate directions. Then each X_k takes one of the $2v$ values $\pm e_1, \dots, \pm e_v$ with equal probability $1/(2v)$. The v components $X_{k,j}$ of X_k hence take values ± 1 with probability $1/(2v)$ apiece and value 0 with probability $1 - 1/v$. Thus, each $X_{k,j}$ has mean zero and variance $1/v$. Arguing as for the walk in the plane, the $X_{k,j}$ are dependent but mutually uncorrelated, $\text{Cov}(X_{k,i}, X_{k,j}) = 0$ if $i \neq j$. Thus, X_k has mean zero and diagonal covariance matrix $\Sigma = \frac{1}{v}I_v$ where I_v is the $v \times v$ identity matrix. Now $\Phi_{\frac{1}{v}I_v}$

¹⁵S. C. Fang and S. S. Venkatesh, "The capacity of majority rule", *Random Structures and Algorithms*, vol. 12, pp. 83–109, 1998.

¹⁶C. Stein, *Approximate Computation of Expectations*, op. cit.

is a product of ν identically distributed marginal normals with mean zero and variance $1/\nu$, in notation, $\Phi_{\frac{1}{\sqrt{\nu}}I_\nu} = \Phi_{0,1/\sqrt{\nu}}^{\otimes \nu}$ where, with the conventions of Section X.1, $\Phi_{0,1/\sqrt{\nu}}$ represents a normal distribution with mean 0 and variance $1/\nu$ in one dimension. It follows then that S_n^* has asymptotically independent normal components each with mean zero and variance $1/\nu$. As an illustration of the simplicities that an asymptotic viewpoint brings, the probability that S_n lies in an axis-parallel rectangle of side $2a\sqrt{n}/\sqrt{\nu}$ centred at the origin tends asymptotically to $[\Phi(a) - \Phi(-a)]^\nu = [2\Phi(a) - 1]^\nu$, which expression is as close to one as desired for a sufficiently large choice of a : after n steps the random walk is essentially confined to a box of side-length of order \sqrt{n} . ▶

If we replace the random walk steps X_k in the previous example by vectors of fixed length but random direction we obtain the discretised random flight model considered in Section IX.3.

EXAMPLE 2) Lord Rayleigh's random flights. Suppose X_1, X_2, \dots are random points on the unit sphere \mathbb{S}^2 in three dimensions. By reasons of symmetry it is clear that the components of each X_k have zero mean, $E(X_{k,j}) = 0$ for each j . As the sum of the squares of the components $X_{k,j}$ is identically one, additivity of expectation shows that the sum of the variances of the $X_{k,j}$ is equal to one and another appeal to symmetry shows that each $X_{k,j}$ has variance $1/3$. A final appeal to symmetry shows that we must also have $E(X_{k,i}X_{k,j}) = 0$ for each pair of distinct indices i and j and it follows that X_k has covariance matrix $\Sigma = \frac{1}{3}I_3$ where I_3 is the 3×3 identity matrix. The components of X_k are again uncorrelated but, as the sum of their squares has to be one, manifestly dependent.

The resultant S_n is the vector sum $X_1 + \dots + X_n$ and is identified as an arrow in three dimensions. While the distribution of the resultant is given in general by a messy convolution, with the proper norming matters simplify dramatically, at least asymptotically: the scaled resultant $S_n^* = S_n/\sqrt{n}$ converges in distribution to the normal $\Phi_{\frac{1}{3}I_3}$ courtesy our friendly neighbourhood central limit theorem. Again, $\Phi_{\frac{1}{3}I_3}$ is a product of three marginal normal distributions, each of mean zero and variance $1/3$ (uncorrelated joint normals are independent). In consequence, the components of S_n^* are asymptotically independent with marginal normal distributions with mean 0 and variance $1/3$. Hence, the squared length of the scaled resultant S_n^* is asymptotically distributed as the sum of the squares of three independent normals and, by the development in Section X.2, is asymptotically governed by the chi-squared density $g_{3/2}(t; 3/2)$. In view of Example X.2.4, the length of S_n^* is hence governed by Maxwell's density

$$2tg_{3/2}(t^2; 3/2) = \frac{3\sqrt{6}}{\sqrt{\pi}} t^2 e^{-3t^2/2} \quad (t > 0)$$

asymptotically. A consideration of the difficulties involved in a frontal attack on the problem brings a renewed appreciation of the simplicity of the asymptotic

argument.

The analysis is unchanged if X_k is identified with a *random direction* in space whose mean squared length is one. Formally, X_k is identified with a product $L_k \Theta_k$ where Θ_k is a random point on the surface of the unit ball and L_k is a positive random variable independent of Θ_k with $E(L_k^2) = 1$. If $\|X_k\|$ represents the (Euclidean) length of X_k then $\|X_k\|^2 = L_k^2$ as Θ_k has length one and it follows that $E(\|X_k\|^2) = E(L_k^2) = 1$. Nothing changes in our analysis now and, in consequence, the scaled resultant S_n^* converges in distribution to $\Phi_{\frac{1}{3}I_3} = \Phi_{0,1/\sqrt{3}}^{\otimes 3}$, the product of three marginal normals with mean zero and variance $1/3$, while the length of S_n^* is governed asymptotically by Maxwell's density given above.

We need not be confined of course to three dimensions. In v dimensions the corresponding analysis shows that the scaled resultant S_n^* converges in distribution to $\Phi_{\frac{1}{v}I_v} = \Phi_{0,1/\sqrt{v}}^{\otimes v}$ so that its components are again asymptotically independent and normal with mean zero and variance $1/v$; the square of the length of the scaled resultant S_n^* is then governed asymptotically by the chi-squared density $g_{v/2}(t; v/2)$ whence the length of S_n^* is governed asymptotically by the v -dimensional analogue of Maxwell's density

$$2tg_{v/2}(t^2; v/2) = \frac{v^{v/2}}{2^{v/2-1}\Gamma(v/2)} t^{v-1} e^{-t^2 v/2} \quad (t > 0).$$

For an analysis in a different spirit see Section IX.3 where we had considered the projection of the resultant along any axis. ►

The lengths of random walks crop up, surprisingly, also in tests for randomness and fraud. In consequence, the multivariate central limit theorem plays an important rôle in these settings as we see in the following sections.

10 A test statistic for aberrant counts

A population has members of v categories (political parties, factions, genera, colours, or classes) in proportions p_1, \dots, p_v . In a random sample of size n drawn with replacement from the population these categories are found represented with counts of N_1, \dots, N_v in each of the categories. Are the observed category frequencies $\frac{1}{n}N_1, \dots, \frac{1}{n}N_v$ in the sample representative of the assumed category proportions p_1, \dots, p_v or are one or more of the observed counts aberrant?

Clearly, a definitive answer cannot obtain for this question. The best we can do is articulate a principled test procedure and check whether the observed counts satisfy the test. What kind of test should one adopt? Again there is no definitive answer but the following criteria articulated by T. W. Körner may serve as a guide.

1. *The test should be easy to apply.*
2. *The method should not ignore certain pieces of data.*

One may embrace the first criterion on pragmatic as well as aesthetic considerations. Reducing computational complexity is always a worthwhile goal and simpler criteria tend to be more persuasive on philosophical grounds following the urgings of the fourteenth-century Franciscan friar and philosopher William of Ockham who advocated parsimony in thought.¹⁷ And the second criterion is certainly unexceptionable; a test that focuses only on the count N_1 , for instance, will be quickly rejected as being not representative enough.

With these criteria in mind let us consider the random variables N_1, \dots, N_v . It is easy to see that the vector of counts $\mathbf{N} = (N_1, \dots, N_v)$ has a multinomial distribution,

$$P\{N_1 = n_1, \dots, N_v = n_v\} = \frac{n!}{n_1! \cdots n_v!} p_1^{n_1} \cdots p_v^{n_v}, \quad (10.1)$$

but it is not so easy to see how to fashion a reasonable test for aberration in v dimensions. For instance, while in principle it is possible to identify a set \mathbb{N} of lattice points (n_1, \dots, n_v) on which the multinomial probabilities (10.1) sum to at least, say, 0.95 (that is to say, we form a 95% confidence interval in v dimensions), a test that rejects the counts as aberrant if \mathbf{N} is not in \mathbb{N} is computationally difficult and fails Körner's first maxim—the region \mathbb{N} , depending critically as it does on the category probabilities p_1, \dots, p_v , does not have a clean characterisation in general, nor is it clear which of the many such regions that can be formed should be adopted. Perhaps a consideration of the individual counts N_j instead of them all collectively will lead to a simpler test.

As a trial results in category j with probability p_j , in a succession of n independent trials the number N_j of occurrences of category j subscribes to a binomial distribution corresponding to n tosses of a coin with success probability p_j . It follows that N_j has mean np_j and variance $np_j(1 - p_j)$. Almost definitely we will want to compare N_j to its mean value as a large deviation from the mean is a certain indication of aberration. We will also want to normalise the deviations from the mean by dividing by a factor proportional to \sqrt{n} to keep the exploding variance under control. We are hence led to consider deviations from the mean of the form $S_{n,j}^* := (N_j - np_j)/\sqrt{n}\alpha_j$. What should the positive normalising factors α_j be chosen as? It is certainly tempting to set $\alpha_j = \sqrt{p_j(1 - p_j)}$ as that would standardise $S_{n,j}^*$ to unit variance, but the dependence across the N_j gives pause and it is wise to give ourselves a little room to manoeuvre.

¹⁷Ockham favoured spare explanations to over-elaborate ones and put the principle of parsimony to striking use across his writing. The phrase “*Frustra fit per plura quod potest fieri per pauciora*” [loosely translated, it is futile to do with more than which can be done with less] appearing in his work *Summa Totius Logicae* written around 1323 CE provides a typical example of his exhortations on the subject. The principle of parsimony has now become ingrained in logical and scientific exploration and is commonly referred to as *Ockham's razor*.

We accordingly obtain a vector of centred and normalised deviations from the mean, $\mathbf{S}_n^* = (S_{n,1}^*, \dots, S_{n,v}^*)$, as a candidate for a test. As positive or negative deviations of counts from their means are equally pernicious, it is natural to consider the square of the components of \mathbf{S}_n^* as a measure of their aberrance and, following Körner's second exhortation to not ignore any piece of data, we accordingly are led to a consideration of the test statistic¹⁸

$$\chi^2 := \sum_{j=1}^v \left(\frac{N_j - np_j}{\sqrt{n} \alpha_j} \right)^2 = \sum_{j=1}^v [S_{n,j}^*]^2 = \|\mathbf{S}_n^*\|^2 \quad (10.2)$$

formed as the square of the Euclidean length of \mathbf{S}_n^* . Excessively large or small values for the statistic signal aberrant values for one or more of the category counts N_j .

This test statistic is certainly simple to compute. To be able to fashion a test out of this procedure, however, we will have to determine the distribution of \mathbf{S}_n^* and understand the rôle played by different selections of the normalising constants α_j . We consider these issues in turn.

Our experience with the de Moivre–Laplace theorem for the binomial suggests that we may be able to find a multidimensional normal approximation for the multinomial (10.1). Accordingly, let us revisit the sequence of trials in the experiment.

The abstract setting is that of throws of a v -sided die whose faces occur with probabilities p_1, \dots, p_v . A throw of the die may be represented by a random vector $\mathbf{Z} = (Z_1, \dots, Z_v)$ precisely one of whose components takes value 1 with the remaining components taking value 0. If e_1, \dots, e_v are the standard basis vectors in \mathbb{R}^v then \mathbf{Z} takes value e_j with probability p_j , the occurrence of the event $\{\mathbf{Z} = e_j\}$ connoting that the result of the throw was the j th face. For each j then, we have $E(Z_j) = p_j$ and $\text{Var}(Z_j) = p_j(1 - p_j)$. Moreover, if $i \neq j$, $\text{Cov}(Z_i, Z_j) = E(Z_i Z_j) - p_i p_j = -p_i p_j$ as precisely one of the components of \mathbf{Z} is non-zero.

In this setting, a random sample of size n is a sequence of independent vectors Z_1, \dots, Z_n with each of the points $Z_k = (Z_{k,1}, \dots, Z_{k,v})$ having the same distribution as \mathbf{Z} . As $N_j = \sum_{k=1}^n Z_{k,j}$, we have

$$S_{n,j}^* = \frac{N_j - np_j}{\sqrt{n} \alpha_j} = \frac{1}{\sqrt{n}} \sum_{k=1}^n \frac{Z_{k,j} - p_j}{\alpha_j}.$$

For each k it is now natural to associate with each Z_k its centred and scaled variant $X_k = (X_{k,1}, \dots, X_{k,v})$ whose components are given by an affine linear shift of the corresponding components of Z_k via $X_{k,j} = (Z_{k,j} - p_j)/\alpha_j$. The sequence of random vectors X_1, \dots, X_n inherits independence from the Z_k , with

¹⁸In this context, the word “statistic” is used in the sense peculiar to the discipline of statistics to mean a function of the data.

the X_k sharing a common distribution. By linearity of expectation, it follows quickly that each of the X_k has mean 0 and covariance matrix $\Sigma = [\sigma_{i,j}]$ where

$$\begin{aligned}\sigma_{j,j} &= \text{Var}(X_j) = p_j(1-p_j)/\alpha_j^2, \\ \sigma_{i,j} &= \text{Cov}(X_i, X_j) = -p_i p_j / \alpha_i \alpha_j \quad (i \neq j).\end{aligned}\tag{10.3}$$

With these definitions in hand, we may write $S_n^* = \frac{1}{\sqrt{n}} \sum_{k=1}^n X_k$ as a sum of independent random vectors and the multivariate central limit theorem beckons. We need to proceed circumspectly, however, as the components of each X_k are linearly dependent. Indeed, for each k ,

$$\alpha_1 X_{k,1} + \cdots + \alpha_v X_{k,v} = (Z_{k,1} + \cdots + Z_{k,v}) - (p_1 + \cdots + p_v) = 0 \tag{10.4}$$

as precisely one of the components of Z_k is equal to one with the others all identically zero. It follows that the vectors X_1, \dots, X_n , hence also the vector S_n^* , all lie in the $(v-1)$ -dimensional subspace $\mathbb{G}_{v-1} := \{x : \alpha_1 x_1 + \cdots + \alpha_v x_v = 0\}$ orthogonal to the vector $\alpha = (\alpha_1, \dots, \alpha_v)$. And *a fortiori* the common covariance matrix Σ of the X_k is degenerate. Indeed, we may write (10.4) succinctly in vector notation as $\alpha X_k^\top = 0$. Postmultiplying both sides by X_k and taking expectation, it follows by linearity of expectation that $0 = E(\alpha X_k^\top X_k) = \alpha E(X_k^\top X_k) = \alpha \Sigma$. As α is not identically zero it follows that the covariance matrix Σ is singular and the central limit theorem must amend before we can utilise it.

How to proceed? On consideration of the geometry of the setting, it becomes irresistible to rotate the coordinate system so that one axis is orthogonal to \mathbb{G}_{v-1} (see Figure 1). We may without loss of generality suppose that α has unit length (as scaling α only results in a scalar multiplication of the test statistic (10.2) by $1/\|\alpha\|$) and we may as well select the v th coordinate axis in the rotated system as the one orthogonal to \mathbb{G}_{v-1} . Accordingly, suppose q_1, \dots, q_{v-1}, q_v are v mutually orthogonal vectors¹⁹ in \mathbb{R}^v , $q_j q_k^\top = \delta_{jk}$, the Kronecker delta function taking value 1 when $i = j$

and value 0 when $j \neq k$, and with $q_v = \alpha$ the unit-length vector orthogonal to \mathbb{G}_{v-1} . The system q_1, \dots, q_v then forms an orthonormal basis for \mathbb{R}^v . We are now naturally led to consider the transformation $X \mapsto XQ$ where Q is the orthogonal matrix whose columns $q_1^\top, \dots, q_{v-1}^\top, q_v^\top = \alpha^\top$ are the transposes of the unit vectors comprising the selected orthonormal basis. The transformation Q merely rotates the coordinate system to align with the orthonormal eigenvector basis. Writing $Q' = (q_1^\top \ q_2^\top \ \dots \ q_{v-1}^\top)$ for the $v \times v-1$ matrix whose

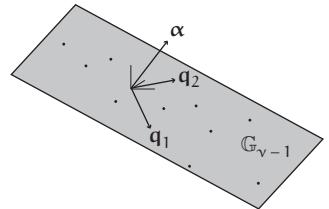


Figure 1: The rotated coordinate frame.

¹⁹The reader should keep in mind the *convention*, introduced in Chapter VII, that in matrix–vector operations, vectors are to be treated as row vectors, their transposes becoming column vectors.

columns span the $(n - 1)$ -dimensional subspace $\mathbb{G}_{\nu-1}$, we have $Q = (Q' \quad q_\nu^T)$. In the rotated coordinate system, each X_k is mapped into a vector whose ν th coordinate is zero, $X_k \mapsto X_k Q = (X'_k, 0)$, where the $(\nu - 1)$ -dimensional X'_k is given by $X_k Q'$. By additivity of expectation, it is clear that X'_k has zero mean, while additivity again shows that the corresponding covariance matrix is given by

$$\Sigma' = \text{Cov}(X'_k) = E(X'^T_k X'_k) = Q'^T E(X'_k X_k) Q' = Q'^T \Sigma Q'.$$

As S_n^* is a linear form of the X_k , in the rotated frame it is also mapped into a vector whose ν th coordinate is zero, $S_n^* \mapsto S_n^* Q_\nu = (S_n^{*\prime}, 0)$, where the $(\nu - 1)$ -dimensional $S_n^{*\prime}$ is given by

$$S_n^{*\prime} = S_n^* Q' = \frac{1}{\sqrt{n}} \sum_{k=1}^n X_k Q' = \frac{1}{\sqrt{n}} \sum_{k=1}^n X'_k.$$

As the vectors X'_1, \dots, X'_n are independent and share a common distribution with mean zero and non-degenerate covariance matrix Σ' , it follows readily via the central limit theorem that $S_n^{*\prime}$ converges in distribution to the normal distribution $\Phi_{\Sigma'}$ in $\nu - 1$ dimensions.

The key here is that our selected statistic χ^2 in (10.2) depends only upon the Euclidean length of S_n^* and, as the reader may recall, rotational transformations leave lengths unchanged. In the present setting,

$$\|S_n^*\|^2 = \|(S_n^{*\prime}, 0)\|^2 = S_n^* Q Q^T S_n^{*\prime T} = S_n^* S_n^{*\prime T} = \|S_n^*\|^2,$$

as the matrix Q is orthogonal. For every choice of positive constants a and b with $a < b$, it follows by (10.2) that

$$P\{a < \chi^2 < b\} = P\{a < \|S_n^*\|^2 < b\} \rightarrow \int_{z:a < \|z\|^2 < b} \phi_{\Sigma'}(z) dz \quad (10.5)$$

and we have a solution, in principle, for our problem though, as a practical matter, the $(\nu - 1)$ -dimensional integral on the right is not computationally pleasant.

The normal density $\phi_{\Sigma'}$ that appears in the integral depends implicitly upon the choice of the scaling parameters $\alpha = (\alpha_1, \dots, \alpha_\nu)$. While any selection of positive α will work, with a proper choice, pouf! the computation becomes virtually transparent. A consideration of the cases $\nu = 2$ and $\nu = 3$ points the way (after, it must be confessed, some very tedious algebra).

Let X be a generic vector in the subspace $\mathbb{G}_{\nu-1}$ with the distribution of the X_k and in the rotated system suppose $X \mapsto XQ = (X', 0)$ where $X' = (X'_1, \dots, X'_{\nu-1})$.

LEMMA *With the choice $\alpha_j = \sqrt{p_j}$ for each j , the random vector X' has uncorrelated components. More precisely, for each j and every pair $i \neq j$, we have $\text{Var}(X'_j) = 1$ and $\text{Cov}(X'_i, X'_j) = 0$, whence the covariance matrix $\Sigma' = \text{Cov}(X')$ is just the identity matrix of order $\nu - 1$.*

PROOF: Let $\mathbf{t} = (t_1, \dots, t_v)$ be any v -dimensional vector in the original coordinate system. In the rotated system then, $\mathbf{t} \mapsto \mathbf{t}Q = (t', t'_v)$, where $t' = (t'_1, \dots, t'_{v-1})$. As the reader is no doubt aware, rotations preserve projections, $\mathbf{X}'\mathbf{t}'^T = (\mathbf{X}', 0)(\mathbf{t}', t'_v)^T = \mathbf{X}\mathbf{Q}\mathbf{Q}^T\mathbf{t}^T = \mathbf{X}\mathbf{t}^T$, and in consequence, $E\{(\mathbf{X}'\mathbf{t}'^T)^2\} = E\{(\mathbf{X}\mathbf{t}^T)^2\}$. These quantities represent the expected value of the square of the projections of \mathbf{X}' and \mathbf{X} onto \mathbf{t}' and \mathbf{t} , respectively, and we compute them in turn. In the rotated frame, the v th coordinate of \mathbf{t}' is given by $t'_v = \mathbf{t}'\boldsymbol{\alpha}^T = \sqrt{p_1}t_1 + \dots + \sqrt{p_v}t_v$. Specialising (10.3) to $\alpha_j = \sqrt{p_j}$, it follows that

$$\begin{aligned} E\{(\mathbf{X}\mathbf{t}^T)^2\} &= \sum_{i=1}^v \sum_{j=1}^v E(X_i X_j) t_i t_j = \sum_{j=1}^v E(X_j^2) t_j^2 + \sum_{i=1}^v \sum_{j \neq i} E(X_i X_j) t_i t_j \\ &= \sum_{j=1}^v (1 - p_j) t_j^2 - \sum_{i=1}^v \sum_{j \neq i} \sqrt{p_i p_j} t_i t_j = \sum_{j=1}^v t_j^2 - \left(\sum_{j=1}^v \sqrt{p_j} t_j \right)^2 = \|\mathbf{t}\|^2 - t'_v^2. \end{aligned}$$

But, as rotations leave length unchanged, we have $\|\mathbf{t}\|^2 = \|(\mathbf{t}', t'_v)\|^2 = t'_1^2 + \dots + t'_{v-1}^2 + t'_v^2$ so that we finally obtain $E\{(\mathbf{X}\mathbf{t}^T)^2\} = \sum_{j=1}^{v-1} t_j'^2$. On the other hand, by simply expanding out the projection of \mathbf{X}' onto \mathbf{t}' , we have

$$E\{(\mathbf{X}'\mathbf{t}'^T)^2\} = E\left\{ \left(\sum_{j=1}^{v-1} X'_j t'_j \right)^2 \right\} = \sum_{i=1}^{v-1} \sum_{j=1}^{v-1} E(X'_i X'_j) t'_i t'_j.$$

As the choice of the vector \mathbf{t} is arbitrary, a term-by-term comparison of the expressions for $E\{(\mathbf{X}'\mathbf{t}'^T)^2\}$ and $E\{(\mathbf{X}\mathbf{t}^T)^2\}$ does the trick. ▶

With the prescient choice $\alpha_j = \sqrt{p_j}$ for the scaling factors, the random variables X'_1, \dots, X'_{v-1} are uncorrelated but they are *not* independent. The power of the proper choice of normalisation is evident only when we consider the effect of the sum of a large number of independent vectors \mathbf{X}_k distributed as \mathbf{X} . The weak independence property manifest in component uncorrelatedness is now translated into a strong “asymptotic independence” for the components of the weighted sum $\mathbf{S}_n^{*'}$. In more detail, with $\alpha_j = \sqrt{p_j}$, the covariance matrix Σ' of $\mathbf{S}_n^{*'}$ is just the identity matrix of order $v - 1$. It follows that its limiting multivariate normal density $\phi_{\Sigma'}$ devolves into the product of $v - 1$ univariate normals, $\phi_{\Sigma'}(\mathbf{z}) = \phi(z_1) \cdots \phi(z_{v-1})$, where, as usual, $\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$ is the standard normal density. It follows that $\mathbf{S}_n^{*'}$ has asymptotically independent normal components and, in consequence, the integral on the right side of (10.5) becomes

$$\int \cdots \int_{a < z_1^2 + \cdots + z_{v-1}^2 < b} \phi(z_1) \cdots \phi(z_{v-1}) dz_1 \cdots dz_{v-1}. \quad (10.6)$$

Let Z_1, \dots, Z_{v-1} be independent, standard normal random variables and let $V^2 = Z_1^2 + \cdots + Z_{v-1}^2$. We may now identify the integral above as the probability

of the event $a < V^2 < b$. But, as we saw in Section X.2, V^2 has the chi-squared density with $\nu - 1$ degrees of freedom given by

$$g_{(\nu-1)/2}(t; 1/2) = \frac{1}{2^{(\nu-1)/2} \Gamma((\nu-1)/2)} t^{(\nu-3)/2} e^{-t/2} \quad (t > 0). \quad (10.7)$$

Introducing the nonce notation $G_{\nu-1}(t) = P\{V^2 \leq t\}$ for the chi-squared d.f., the integral in (10.6) may be identified with $G_{\nu-1}(b) - G_{\nu-1}(a)$, proving the

THEOREM *The statistic $\chi^2 = \sum_{j=1}^{\nu} (N_j - np_j)^2 / np_j$ has an asymptotic chi-squared distribution with $\nu - 1$ degrees of freedom. In particular, if $0 \leq a < b$, then*

$$P\{a < \chi^2 < b\} \rightarrow G_{\nu-1}(b) - G_{\nu-1}(a) \quad (n \rightarrow \infty).$$

This elegant theorem was discovered by Karl Pearson, the “founder of modern statistics”, and has had an abiding impact on statistics.²⁰ The chi-squared test, together with the Kolmogorov–Smirnov test for densities, forms a staple in the testing of randomness.

11 A chi-squared test

The chi-squared statistic provides a simple vehicle for querying whether population data truly fit a presupposed random model. Suppose data are drawn by independent sampling from an underlying distribution (population) with elements classified in ν categories. It is assumed *a priori* that these categories arise on sampling with probabilities p_1, \dots, p_ν . In the language of statistics, the assumed model probabilities p_1, \dots, p_ν form the *null hypothesis* (or default state) of nature. Suppose a random sample of size n engenders category counts N_1, \dots, N_ν . The chi-squared test statistic $\chi^2 = \sum_{j=1}^{\nu} (N_j - np_j)^2 / (np_j)$ now provides an exquisitely simple platform for querying whether the data are consistent with the assumed model probabilities. Aberration in the counts N_j is manifest in an improbably large or small value for the statistic which then provides a principled basis on which to either accept the assumed model probabilities as being consistent with the observations or reject them as being unlikely to have generated the observed data. The desired level of confidence in the test is captured by the specification of a *confidence interval* $[a, b]$ for the statistic χ^2 : *we accept the category counts N_j as fluctuating within reasonable ranges from their mean values if $a \leq \chi^2 \leq b$, and reject the counts as anomalous if $\chi^2 > b$ or $\chi^2 < a$.* The model probabilities are rejected when $\chi^2 > b$ on the grounds that one or more of the N_j deviate too far from their mean values to be attributable merely

²⁰K. Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling”, *Philosophical Magazine*, Series 5, vol. 50 (302), pp. 157–175, 1900.

to chance; the rejection of the model when $\chi^2 < a$, on the other hand, is because of too excessive a regularity to be expected from chance.

How then should one select the confidence interval $[a, b]$? A too stringent definition of aberration will result in many entirely reasonable chance fluctuations being rejected, while a very lax definition will admit spurious cases. A compromise is effected by selecting an interval $[a, b]$ so that the event that χ^2 is outside this interval has a sufficiently low probability δ . A conventional choice is $\delta = 0.05$ though any $\delta > 0$ may be specified, the choice depending on the needs of the application. Thus, $\chi^2 \notin [a, b]$ sufficiently rarely that, if this occurs, we can conclude with high confidence $1 - \delta$ (say, 95%) that the data are indeed aberrant for the assumed model. It only remains to determine the *quantiles* a and b for a desired level of confidence $1 - \delta$.

Surely, excessively large values of the statistic χ^2 are suspicious but the reader may wonder why we would cavil at small values, a feeling that is shared by many commercial makers of statistical software. The statistical analysis package on my computer, in common with most commercially available statistical routines, routinely provides 95% confidence intervals $[0, b]$ for chi-squared tests by computing the quantile $b = G_{\nu-1}^{-1}(0.95)$ for which the area under the right tail (b, ∞) of the chi-squared density with $\nu - 1$ degrees of freedom equals 5%. If the computed statistic χ^2 exceeds b then the data are deemed aberrant; else not. Thus, large values for χ^2 are ruled aberrant but not small values.

Is there a reason to be suspicious of too small a value for the statistic χ^2 ? A consideration of the density (10.7) shows that for $\nu = 2$ it is unbounded at zero and decreases monotonically, for $\nu = 3$ it decreases monotonically from a maximum value of $1/2$ at the origin, and for $\nu \geq 4$ it increases to a unique maximum at $y = \nu - 3$ and decreases monotonically thereafter. If the reader does not feel like taking derivatives to verify the statement she may find the graphs in Figure 2 persuasive.

For $\nu = 2$ and 3, the chosen confidence interval should clearly have zero as a left endpoint—the region around the origin has the largest probability. But for $\nu \geq 4$, it is very unlikely that the chi-squared value will land in the immediate vicinity of the origin and a χ^2 value too close to zero indicates a rather too suspicious regularity in the data. In these cases, if one is agnostic about the nature of aberration, either too variable or not variable enough, one should elect a confidence interval $[a, b]$ excluding the region around the origin. An appeal to Ockham's razor may suggest that a principled choice of interval would be to select the *smallest* interval of confidence $1 - \delta$ (nominally, 95%) as providing the tightest range of non-aberrant values for the statistic for the desired level of confidence. A lack of symmetry in the density about its maximum value makes the computation of the smallest interval non-trivial and in a nod to computational convenience we may adopt the heuristic of selecting the interval $[a, b]$ so that both the left and right tails, $[0, a]$ and (b, ∞) , have equal probability. In

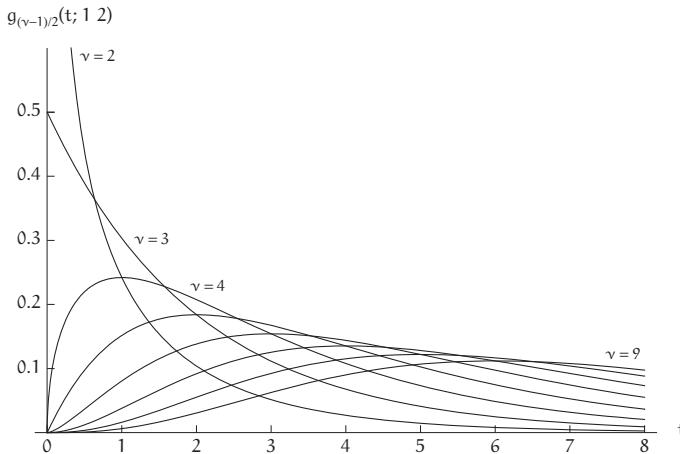


Figure 2: Chi-squared densities with $\nu - 1$ degrees of freedom.

other words, for a confidence level of $1 - \delta$, we select a and b as the quantiles $G_{\nu-1}^{-1}(\delta/2)$ and $G_{\nu-1}^{-1}(1 - \delta/2)$, respectively.

EXAMPLES: 1) *Testing random number generators.* A variety of computer algorithms and internet protocols have randomness baked into the cake as it were. For instance, the ALOHA protocol governs how users who share a communication medium handle collisions when two or more users wish to use the medium at the same time. The protocol mandates that each colliding user back off by a random amount of time, the parameters chosen to minimise the overall delay while ensuring that repeat collisions are very unlikely. Another instance of the use of randomness is in crafting defences against denial of service (DoS) attacks over the internet in which an adversary attempts to overwhelm a user's computational resources by bombarding her with spurious packets. The attack can be defanged by an appropriately chosen randomised selection strategy which examines only a small fraction of incoming packets.²¹ In cases such as these it becomes essential to have access to a good random number generator.

Table 1 shows sampling data for four generators, (A), (B), (C), and (D) producing integers purportedly at random in the range 1 through 10. A sample of 100 digits was generated using each of the generators, and the sample counts for each digit shown. On eyeballing the data, generator (A) may seem to be rather too violent in its fluctuations (especially for digits 9 and 10), generator (B), on the other hand, may appear too good to be true, while (C) and (D) are somewhat more equivocal. The stage is set for a chi-squared test with 9 degrees of freedom for equal digit probabilities $p_j = 1/10$ ($1 \leq j \leq 10$) and a sample of

²¹C. Gunter *et al.*, *op. cit.*

Normal Approximation

Digits j	1	2	3	4	5	6	7	8	9	10	χ^2
Digit counts (A)	13	6	9	9	8	2	6	3	1	43	133.0
Digit counts (B)	11	9	9	9	12	10	10	9	10	11	1.0
Digit counts (C)	11	12	15	9	9	9	8	10	8	9	4.2
Digit counts (D)	9	9	6	13	9	12	6	13	9	14	7.4

Table 1: Data from four pseudo-random number generators.

size $n = 100$. The statistical analysis package on my computer gives the quantiles $a = G_9^{-1}(0.025) = 2.7$ and $b = G_9^{-1}(0.975) = 19.0$ (both figures rounded off) for a 95% confidence interval $[2.7, 19.0]$ with equal tail probabilities of 2.5% on either side. The corresponding values of the chi-squared test statistic for the four generators are provided in the last column of the table.

The chi-squared test with acceptance region $2.7 \leq \chi^2 \leq 19.0$ rejects generator (A) as being excessively irregular, rejects generator (B) on the grounds of a too suspicious regularity (perhaps it's broken), and accepts generators (C) and (D) as producing variability within acceptable bounds. In this context, it is better for the chi-squared value to be slightly higher than lower. And it is as well that generators (C) and (D) were the ones on offer in the statistical package on my computer.

Incidentally, the selfsame statistical package on my computer is inclined to accept the bona fides of generator (B) as it uses as default the confidence interval $[0, 16.9]$ which places all the aberration probability of 5% in the right tail. For the reasons I have given above, I would reject this generator on the basis that a two-sided confidence interval provides a truer picture of aberration as well as a (slightly, in this case) more compact interval.

2) A random walk in Wall Street.²² How well does a market trader or financial analyst do in predicting stock movement? In Table 2, data are presented for 125 analysts each of whom picks four stocks from the New York Stock Exchange to outperform the S&P 500 Index over a 90-day period. The number of stocks

Success categories j	0	1	2	3	4
Analyst frequencies N_j	3	23	51	39	9
Random walk frequencies $n p_j$	7.8125	15.6250	46.8750	15.6250	7.8125

Table 2: Analyst performance vis à vis the S&P 500.

picked by a given analyst that outperform the S&P 500 Index is an integer be-

²²This example is taken from D. K. Hildebrand, R. L. Ott, and J. B. Gray, *Basic Statistical Ideas for Managers*, Second Edition, p. 393. Belmont, CA: Thomson Brooks/Cole, 2005.

tween 0 and 4 and the sample of 125 analysts is spread across the five success categories from 0 through 4.

How can we determine whether these data are indicative of analyst acumen? In a *random walk model* for stocks, the rise or fall of any given stock over a given period is assumed to be a symmetric Bernoulli trial. If the analysts were doing no better than the random walk model then the number of stock successes for a given analyst is given by the binomial distribution, $p_j = b_4(j; 1/2) = \binom{4}{j} 2^{-4}$ ($0 \leq j \leq 4$). The expected frequencies for a sample size n are shown tabulated above.

The setting is ripe for a chi-squared test with 4 degrees of freedom, binomial probabilities $p_1 = p_4 = 0.0625$, $p_1 = p_3 = 0.25$, $p_2 = 0.375$, and sample size $n = 125$. My computer gives the chi-squared quantiles for a two-sided test of $a = G_4^{-1}(0.025) = 0.48$ and $b = G_4^{-1}(0.975) = 11.14$ so that the acceptance region for the chi-squared statistic is $[0.48, 11.14]$. The corresponding analyst chi-squared value is $\chi^2 = 7.6$ which is squarely in the acceptance region. Based on the chi-squared analysis alone, one would be forced to conclude that analyst performance is not distinguishable from random guessing. A more nuanced analysis in this case may indicate that there is a slight edge for the analyst. But our chi-squared analysis is not refined enough to capture it. ►

Most uses of the chi-squared test in practice look for large excursions (explaining to some extent the focus of commercial software) but there are uses at the other end of the spectrum as well. As we have seen, a rather too small value for a chi-squared statistic may indicate an honestly malfunctioning system. But there is another possibility. A suspiciously small value may signal the work of an incompetent malign agent, in other words, a bungling fraud, who has cooked the books to reach a desired conclusion. The twenty-first century has provided several cautionary examples of fraud in high finance. In academics the stakes are less consequential (by some measures) but the impact of fraud is no less damaging. An illustration of the chi-squared test put to work in such a context is furnished in the next section.

12 The strange case of Sir Cyril Burt, psychologist

The twentieth-century psychologist Cyril Burt had reached a position of undisputed eminence when, in the year 1946, he became the first psychologist to be knighted. He retired from his academic position at University College, London in 1950 but continued to be very active, publishing widely after retirement till his death in 1971.

Burt had decided views on the heritability of intelligence and published several influential papers on the matter. The data in Table 3 are reproduced

Normal Approximation

TABLE I. DISTRIBUTION OF INTELLIGENCE ACCORDING TO OCCUPATIONAL CLASS: ADULTS

	50– 60	60– 70	70– 80	80– 90	90– 100	100– 110	110– 120	120– 130	130– 140	140+ Total	Mean I.Q.
I. Higher Professional									2	1	3
II. Lower Professional								13	15	1	31
III. Clerical				1	8	16	56	38	3		115.9
IV. Skilled			2	11	51	101	78	14	1		108.2
V. Semiskilled	5	15	31	135	120	17	2			325	97.8
VI. Unskilled	1	18	52	117	53	11	9			261	84.9
Total	1	23	69	160	247	248	162	67	21	2	1000
											100.0

TABLE II. DISTRIBUTION OF INTELLIGENCE ACCORDING TO OCCUPATIONAL CLASS: CHILDREN

	50– 60	60– 70	70– 80	80– 90	90– 100	100– 110	110– 120	120– 130	130– 140	140+ Total	Mean I.Q.
I. Higher Professional						1		1	1		3
II. Lower Professional				1	2	6	12	8	2		31
III. Clerical			3	8	21	31	35	18	6		114.7
IV. Skilled		1	12	33	53	70	59	22	7	1	122
V. Semiskilled	1	6	23	55	99	85	38	13	5		107.8
VI. Unskilled	1	15	32	62	75	54	16	6			258
Total	2	22	70	159	250	247	160	68	21	1	1000
											100.0

Table 3: Burt's Tables I and II. Limitations in my understanding of the mathematical typesetting system L^AT_EX have resulted in the tables being very slightly reformatted but they have not otherwise been altered.

from his much-cited 1961 paper²³ and show the distribution of IQs found in Burt's population for fathers and their children in six occupational groups ranging from Higher Professional to Unskilled. (Children were classified according to the occupational class of their parent.) A reader who examines the data may spot a couple of oddities: the sample size is exactly 1000 and the mean IQ reported for both parents and children is exactly 100. This is partially explained by Burt in his paper.

"The frequencies inserted in the various rows and columns were proportional frequencies and in no way represent the number actually examined: from Class I the number examined was nearer a hundred and twenty than three. To obtain the figures to be inserted (number per mille) we weighted the actual numbers so that the proportions in each class should be equal to the estimated proportions for the total population. Finally, for the purposes of the present analysis we have rescaled our assessments of intelligence, so that the mean of the whole group is 100 and the standard deviation 15."

Based on the first two sentences, some commentators have assumed that the proportional scaling of the number of people in each class was the same so that the total size of the sample is $1000 \times 120/3$ or 40,000 adults. But Burt does not actually say this. And, in fact, Burt says remarkably little about the experimental

²³The data reported in Table 3 and the text extract that follows it are reproduced with permission of John Wiley & Sons, Inc.: C. Burt, "Intelligence and social mobility", *British Journal of Statistical Psychology*, vol. 14, pp. 3–24, 1961. Copyright © John Wiley & Sons, 1961.

procedure or the conditions, or provide much detail about anything. According to Burt,

"The surveys and the subsequent inquiries were carried out at intervals over a period of nearly fifty years, namely from 1913 onwards."

And, again from Burt,

"... for the children the bulk of the data was obtained from surveys carried out from time to time in a London borough selected as typical of the whole county."

And, once more from Burt, his occupational classifications were based

"not on prestige or income, but rather on the degree of ability required for the work."

T. W. Körner has provided a wonderfully readable account of the Burt saga in his book *Fourier Analysis* (Cambridge University Press, 1988) and I cannot resist the urge to quote from it. Körner lists the following four violations of engagement in Burt's reporting of experimental data.

1. *Always give details of the experimental procedure (so that people can redo the experiment or, in any case, form their own opinion on the magnitude of possible errors).*
2. *As far as possible give the data in untreated form (so that other people may do their own data analysis).*
3. *When transforming data explain the method used (so that other people may decide for themselves whether your method is appropriate).*
4. *Do not attempt to conceal the weaknesses of your arguments or to ignore the strong points of your opponents.*

(Though, as Körner points out, the last is beyond the moral strength of most of us.) Certainly, based on the presentation of data, charges of sloppiness may be laid at Burt's door but Burt was to be accused of a much more serious charge: fraud.

Shortly after Burt's death, a book published by Leon Kamin in 1974 titled *The Science and Politics of IQ* made the sensational claim that Burt had falsified much of his data, sparking an academic *cause célèbre* with broadsides from supporters and accusers continuing to this day. While several different analyses could be undertaken of his data, let us see what the chi-squared test can say about it. I will follow Körner's account.

Anthropological models (proposed, among others, by Burt) suggest that inheritable characteristics like height, weight, and intelligence (as reflected in IQ tests) should be normally distributed. (The reader may wish to look back

at Galton's height data in Section VII.8 through this prism.) If we were to assume that this were indeed the case for IQ then it is natural to test Burt's data against the normal distribution with mean 100 and variance $15^2 = 225$. Table 4 contains a compaction of the totals from Burt's tables of parents and children in eight IQ ranges where I have combined Burt's two lowest IQ ranges to form the category 70– and, likewise, combined Burt's two highest IQ ranges to form the category 130+. (As T. W. Körner points out, the reader should make up her

IQ range	70-	70-80	80-90	90-100	100-110	110-120	120-130	130+
Parent totals $N_{p,j}$	24	69	160	247	248	162	67	23
Children totals $N_{c,j}$	24	70	159	250	247	160	68	22
Bin probabilities p_j	0.023	0.068	0.161	0.248	0.248	0.161	0.068	0.023

Table 4: Parent and children IQ bin frequencies from Burt's tables with the two lowest and highest IQ bins combined. The normal bin probabilities have been rounded up.

own mind as to whether this is a reasonable procedure.) The table also lists the normal probabilities p_j for each of these bins under an $\mathcal{N}(100, 225)$ distribution.

In attempting to run a goodness of (normal) fit on these data we immediately run into the problem that we don't know what the sample size used by Burt really was. Let us suppose, to get forrader, that Burt used a proportional scaling factor r for each IQ class. Then the sample size in use is $n = 1000r$ with each of the class frequency numbers scaled to $rN_{p,j}$ and $rN_{c,j}$ for parents and children, respectively. The chi-squared statistics for parents and children can now be computed from Table 4 in terms of r as

$$\chi_p^2(r) = \sum_{j=1}^8 \frac{(rN_{p,j} - 1000rp_j)^2}{1000rp_j} = 0.13r, \quad \chi_c^2(r) = \sum_{j=1}^8 \frac{(rN_{c,j} - 1000rp_j)^2}{1000rp_j} = 0.20r.$$

Now the quantiles of a chi-squared distribution with 7 degrees of freedom for a 95% confidence interval $[a, b]$ with symmetric tails of 2.5% on each side are given by $a = G_7^{-1}(0.025) = 1.69$ and $b = G_7^{-1}(0.975) = 16.01$. If r were really 40 then the chi-squared values for parents and children clock in at 5.2 and 8.0, respectively, well within the confidence interval $[1.69, 16.01]$ and the fit to the $\mathcal{N}(100, 225)$ distribution is completely unexceptionable. The reader may wish to consider, however, the practicality of administering IQ tests (to parents and children) at an average clip of 1600 per year (or in excess of four IQ tests per day) for fifty years. If, on the other hand, $r = 1$ (or perhaps even $r = 0$ as sceptical commentators like Kamin have suggested), then the chi-squared values are $\chi_p^2 = 0.13$ and $\chi_c^2 = 0.2$, and the fact that both are so small leads to a suspicion that the too excellent fit to the normal has been cooked up. Abstract thought cannot lead us any further and we must look to additional data for inspiration.

My objective here was not so much to examine critically all the arguments that have been made about Burt's data with a view to either exonerating

or convicting him, as to show how in venues such as this there is the opportunity for a principled analysis using statistical tests. Though, as we see in the uncertainties of the previous paragraph, certainty in real-world settings may be too much to ask for. To quote Körner again, “battalions of facts, like battalions of men may not always be as strong as supposed”. The interested reader will readily find a wealth of references on both sides of the argument for Burt. L. Hearnshaw’s critically acclaimed biography²⁴ is a good starting point.

13 Problems

1. Repeated convolutions. Starting with the function $f(x)$ which takes value 1 in the interval $(-1/2, 1/2)$ and 0 elsewhere, recursively define the sequence of functions $\{f_n, n \geq 1\}$ as follows: set $f_1 = f$ and, for each $n \geq 2$, set $f_n = f_{n-1} * f$. Define the functions G_n via $G_n(x) = \int_x^\infty f_n(t) dt$. Estimate $G_n(0)$, $G_n(\sqrt{n})$, and $G_n(n/4)$.

2. The Poisson distribution with large mean. Suppose S_λ has the Poisson distribution with mean λ . Show that $(S_\lambda - \lambda)/\sqrt{\lambda}$ converges in distribution to the standard normal. (*Hint:* The stability of the Poisson under convolution.)

3. Continuation. Let $G_\lambda(t)$ be the d.f. of S_λ/λ . Determine $\lim_{\lambda \rightarrow \infty} G_\lambda(t)$ if (i) $t > 1$, and (ii) $t < 1$. What can you say if $t = 1$?

4. Records. Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables with a common distribution. We say that a *record* occurs at time k if $\max\{X_1, \dots, X_{k-1}\} < X_k$. Prove a central limit theorem for the number R_n of records up to time n .

5. Inversions. Let $\Pi: (1, \dots, n) \mapsto (\Pi_1, \dots, \Pi_n)$ be a random permutation. For each k , let $X_k^{(n)}$ be the number of smaller elements, i.e., 1 through $k-1$, to the right of k in the permutation. Prove a central limit theorem for $S_n = X_1^{(n)} + \dots + X_n^{(n)}$, the total number of *inversions*.

6. Khinchin's weak law of large numbers states that if $\{X_k, k \geq 1\}$ is an independent, identically distributed sequence of random variables with finite mean μ then, with $S_n = X_1 + \dots + X_n$, for every $\epsilon > 0$, we have $P\{|\frac{1}{n}S_n - \mu| > \epsilon\} \rightarrow 0$ as $n \rightarrow \infty$. Prove this using operator methods.

7. Cauchy distribution. Show that the Cauchy distribution is stable and determine its characteristic exponent. More verbosely, if X has a Cauchy distribution, and X_1, \dots, X_n are independent copies of X , then $X_1 + \dots + X_n$ has the same distribution as $n^{1/\alpha}X$ for some characteristic exponent $0 < \alpha \leq 2$.

8. Pareto-type distribution. Suppose that X_1, X_2, \dots are independent with common density $f(x) = c/(|x|^3 (\log|x|)^3)$ for $|x| > 2$ (and zero otherwise). Show that these variables do not satisfy the Liapounov condition but do satisfy the Lindeberg condition.

9. A central limit theorem for runs. As in Section VIII.8, let $R_n = R_n(\omega)$ denote the length of the run of 0s starting at the n th place in the dyadic expansion of a point ω drawn at random from the unit interval. Show that $\sum_{k=1}^n R_k$ is approximately normally distributed with mean n and variance $6n$.

²⁴L. Hearnshaw, *Cyril Burt, Psychologist*. London: Hodder and Stoughton, 1981.

10. Suppose X_1, X_2, \dots is a sequence of independent random variables. For each k , the variable X_k is concentrated at four points, taking values ± 1 with probability $\frac{1}{2}(1 - k^{-2})$ and values $\pm k$ with probability $\frac{1}{2}k^{-2}$. By an easy truncation prove that S_n/\sqrt{n} behaves asymptotically in the same way as if $X_k = \pm 1$ with probability $1/2$. Thus, the distribution of S_n/\sqrt{n} tends to the standard normal but $\text{Var}(S_n/\sqrt{n}) \rightarrow 2$.

11. Continuation. Construct variants of the previous problem where $E(X_k^2) = \infty$ and yet the distribution of S_n/\sqrt{n} tends to the standard normal Φ .

12. Once more, Stirling's formula. For every real number x define its negative part by $x^- = -x$ if $x \leq 0$ and $x^- = 0$ if $x > 0$. Let $\{X_k, k \geq 1\}$ be a sequence of independent random variables drawn from the common Poisson distribution with mean 1 and let $\{S_n, n \geq 1\}$ be the corresponding sequence of partial sums. Write $T_n = (S_n - n)/\sqrt{n}$. Show that $E(T_n^-) = n^{n+1/2}e^{-n}/n!$.

13. Continuation, convergence in distribution. Suppose Z is a standard normal random variable. Show that $T_n^- \xrightarrow{d} Z^-$.

14. Continuation, convergence of expectations. Now, here's the tricky part. Show that $E(T_n^-) \rightarrow E(Z^-) = (2\pi)^{-1/2}$ and pool answers to rediscover Stirling's formula. [Hint: Modify the equivalence theorem.]

15. Central limit theorem for exchangeable variables.²⁵ Consider a family of distributions F_θ indexed by a real parameter θ . We suppose that, for each θ , the d.f. F_θ has mean zero and variance $\sigma^2(\theta)$. A random value Θ is first chosen according to a distribution G and a sequence of variables X_1, X_2, \dots is then drawn by independent sampling from F_Θ . Write $a^2 = E(\sigma^2(\Theta))$. Show that the normalised variable $S_n/(a\sqrt{n})$ converges in distribution to the limiting distribution given by $\int_{-\infty}^{\infty} \Phi\left(\frac{ax}{\sigma(\theta)}\right) dG(\theta)$. The limiting distribution is not normal unless G is concentrated at one point.

16. Galton's height data. Table 1 of Section VII.8 provides Galton's height data. Test the data against the hypothesis that they arise from a normal distribution.

17. Rotations. Suppose $\mathbf{X} = (X_1, \dots, X_n)$ has independent $N(0, 1)$ components. If H is any orthogonal $n \times n$ matrix show that $\mathbf{X}H$ has the same distribution as \mathbf{X} .

18. Continuation, the uniform distribution on the sphere. Deduce that $\mathbf{X}/\|\mathbf{X}\|$ has the uniform law λ_{n-1} on the Borel sets of the unit sphere $\mathbb{S}^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ equipped with the geodesic distance. More specifically, $\lambda_{n-1}(\mathbb{A}) = \lambda_{n-1}(\mathbb{A}H)$ for every orthogonal $n \times n$ matrix H and every Borel set \mathbb{A} in $\mathcal{B}(\mathbb{S}^{n-1})$.

19. Continuation, the infinite-dimensional sphere. Let $\mathbf{R}^{(n)} = (R_1^{(n)}, \dots, R_n^{(n)})$ be a point chosen in \mathbb{S}^{n-1} according to the distribution λ_{n-1} . Show that $P\{\sqrt{n}R_1^{(n)} \leq t\} \rightarrow \Phi(t)$ and $P\{\sqrt{n}R_1^{(n)} \leq t_1, \sqrt{n}R_2^{(n)} \leq t_2\} \rightarrow \Phi(t_1)\Phi(t_2)$. This beautiful theorem connecting the normal to the infinite-dimensional sphere is due to E. Borel.²⁶ This famous paper has sparked a huge literature and is important for Brownian motion and Fock space constructions in quantum mechanics. [Hint: Use Problems X.20 and XVII.5.]

²⁵J. R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, "Central limit theorems for interchangeable processes", *Canadian Journal of Mathematics*, vol. 10, pp. 222–229, 1958.

²⁶E. Borel, "Sur les principes de la theorie cinétique des gaz", *Annales Scientifique l'École Normale Supérieure*, vol. 23, pp. 9–32, 1906.

Part C

APPENDIX

Sequences, Functions, Spaces

1 Sequences of real numbers

c 1-3

To save on repetition, it is convenient to collect in one place some elementary facts about the real numbers and establish some notation. The reader who has not been exposed to these ideas should simply glance at the results (which are quite plausible in their own right) and move on, returning to the proofs included at the end of the section when she has leisure.

In the following, $\{x_n, n \geq 1\}$ (or simply $\{x_n\}$ with n understood to range over the natural numbers) denotes a sequence of real numbers. We say that a sequence $\{x_n\}$ is *increasing* if $x_n \leq x_{n+1}$ for all n and *decreasing* if $x_n \geq x_{n+1}$ for all n . If there is a number M such that $x_n \leq M$ for all n then we say that $\{x_n\}$ is *bounded from above*; likewise, if $x_n \geq M$ for all n then $\{x_n\}$ is *bounded from below*. The sequence $\{x_n\}$ is *bounded* if it is simultaneously bounded from above and below.

While the reader will be familiar with the usual axioms of addition, multiplication, and ordering of real numbers, it would not be amiss to state the basic

COMPLETENESS AXIOM *If $\{x_n\}$ is an increasing sequence which is bounded above, then $\{x_n\}$ converges to some number x .*

The necessity of this axiom is made plausible by considering the increasing sequence of fractions $3, 3\frac{1}{10}, 3\frac{14}{100}, 3\frac{141}{1000}, 3\frac{1415}{10000}, \dots$ converging to the non-fractional number π .¹ The axiom is couched in terms of the intuitive idea of convergent sequences, an idea which the historical record shows took a long time to make precise. Here it is.

DEFINITION 1 We say that $\{x_n\}$ *converges to x* if, for every $\epsilon > 0$, there exists an integer $N = N(\epsilon)$ such that $|x_n - x| < \epsilon$ for all integers $n \geq N$.

Colloquially speaking, x_n gets as close as desired to x *eventually*.

¹While it is part of folklore that π is irrational, the knowledgable reader may quibble that it is not at all trivial to show that it is. True. This objection will lose its force if she replaces π by $\sqrt{2}$ (whose irrationality is easy to show as every school child knows) and the given sequence by the successive truncations of the decimal expansion of $\sqrt{2} = 1.4142\dots$.

If $\{x_n\}$ converges to x we write $x_n \rightarrow x$ as $n \rightarrow \infty$ or $\lim_{n \rightarrow \infty} x_n = x$ or even simply $\lim x_n = x$ when the asymptotic parameter n is clear from context. If $x_n \rightarrow x$ and $x_n \leq x$ for each n , we write $x_n \uparrow x$ to graphically indicate that the convergence is from below; likewise, if $x_n \rightarrow x$ and $x_n \geq x$ for each n , we write $x_n \downarrow x$ to indicate that the convergence is from above.

DEFINITION 2 We say that the sequence $\{x_n\}$ is *Cauchy* if, for every $\epsilon > 0$, there exists an integer $N = N(\epsilon)$ such that $|x_m - x_n| < \epsilon$ for all $m, n \geq N$.

It is clear that if a sequence is convergent then its members must eventually all be close to each other: every convergent sequence is Cauchy. A key consequence of the completeness axiom is that the converse is also true: every Cauchy sequence is convergent. This equivalence frequently provides a simple test for convergence when a limiting value is not clearly in view.

The idea of convergence is baked into the cake of the calculus. The reader will recall for instance that the definition of a derivative requires the differential in the denominator to go to zero and yet requires that a ratio of differentials (fluxions in Newton's terminology) converges to a limiting value. While both our founders, Newton and Leibniz, had a keen intuition of what they wanted from the revolutionary objects they were constructing, neither managed to articulate a clean, unambiguous definition of these objects. Indeed, a glance at the historical record shows how remarkably circumspect their attempts were at describing the fundamental objects of their creation. Here is what Newton had to say on the subject of his fluxions:

“[The ultimate ratio] ... is to be understood as the ratio of the quantities, not before they vanish, nor after, but that with which they vanish.”

And Leibniz's attempts at describing his “infinitely small quantities”, which we today may call differentials, while a little more wordy, were hardly better:

“It will be sufficient if, when we speak of ... infinitely small quantities (i.e., the very least of those within our knowledge), it is understood that we mean quantities that are ... indefinitely small If anyone wishes to understand these as the ultimate things ..., it can be done [in other words, just trust me] ..., ay even though he think that such things are utterly impossible; it will be sufficient simply to make use of them as a tool that has advantages for the purpose of calculation, just as the algebraists retain imaginary roots with great profit.”

The reader will be struck by how uncertain the two founders of the calculus were in describing an object so central to the subject. Indeed, their attempts at description dance around the object and ultimately seem to just give up the whole thing as a bad job and simply appeal to our intuition (or better, theirs). Opponents took gleeful note and piled on the derision. Bishop George Berkeley who wrote the pithy *The Analyst, or a Discourse Addressed to an Infidel Mathematician* aimed some wonderfully barbed comments at the mathematicians who were so dismissive of faith-based theology:

“All these points [of mathematics], I say, are supposed and believed by certain rigorous exactors of evidence in religion, men who pretend to believe no further than they can see But he who can digest a second or third fluxion, a second or third differential, need not, methinks, be squeamish about any point in divinity ... And what are these fluxions? The velocities of evanescent increments. And what are these same evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities?”

After all, if one can swallow on faith a fluxion then one can hardly object to another accepting, say, holy writ, on faith. [The context was new at the time but the wry observation was already part of canonical lore as the reader versed in scripture will recall: “Ye blind guides, who strain out a gnat and swallow a camel!” — Matthew 23:24.]

In spite of its great success, matters were hence still unsettled at the heart of the calculus and it was only a century and a half later in the hugely influential *Cours d'Analyse* (1821) that the great Augustin Louis Cauchy provided a formal definition that avoided the logical pitfalls that had bedevilled Newton and Leibniz. Cauchy managed to avoid the problems of what happens *at the limit* by focusing on the approach to it:

“When the values successively attributed to a particular variable approach indefinitely a fixed value, so as to end by differing from it by as little as one wishes, this latter is called the limit of the others.”

The reader will recall the paradox of the rabbit which, starting one metre from a luscious piece of lettuce, in each step moves one-half of the distance towards it but yet never gets to its destination. The series $1/2 + (1/2)^2 + (1/2)^3 + \dots$ converges to the limit 1 (but not in a finite number of steps). Of course, the hungry rabbit resolves this paradox by moving, not in steps per unit time, but at a constant speed $v = \lim_{\Delta t \rightarrow 0} \Delta x / \Delta t$ or, in more familiar notation, $v = dx/dt$, to its destination.

Cauchy's view of limits still had the deficiency, from our perspective, of conceptually invoking the idea of motion towards the limit. It is this last remaining barrier towards a fully formal, essentially *static* theory that the abstract formulation of Weierstrass allays.

Let \mathbb{A} be any subset of the real line. A number u is an *upper bound* of \mathbb{A} if $x \leq u$ for every $x \in \mathbb{A}$. Any set \mathbb{A} with an upper bound is said to be *bounded from above*. Of course, \mathbb{A} need not have an upper bound at all in which case it is said to be *unbounded*; consider, for instance, the set of all rational numbers.

A number v is called the *least upper bound* of \mathbb{A} if v is an upper bound of \mathbb{A} and v is less than or equal to every other upper bound of \mathbb{A} . The least upper bound is also called the *supremum* of \mathbb{A} and denoted $\sup \mathbb{A}$. It is easy to see that if \mathbb{A} has a least upper bound then it must be unique though it need not lie in \mathbb{A} ; if, for instance, \mathbb{A} is the open interval $(0, 1) = \{x : 0 < x < 1\}$ then $\sup \mathbb{A} = 1$ but 1 is not an element of \mathbb{A} . If \mathbb{A} is unbounded above there is no harm in writing $\sup \mathbb{A} = +\infty$.

THEOREM 1 *A real number v is the least upper bound of \mathbb{A} if, and only if, v is an upper bound of \mathbb{A} and, for every $\epsilon > 0$, there exists $x \in \mathbb{A}$ such that $x > v - \epsilon$.*

This is a simple but important observation. In other words, there are points in \mathbb{A} arbitrarily close to, and below, v ; or, in yet other words, $v = \sup \mathbb{A}$ if, and only if, v is an upper bound of \mathbb{A} and there is an increasing sequence of numbers $x_n \in \mathbb{A}$ such that $x_n \rightarrow v$ as $n \rightarrow \infty$. The proof is not at all difficult and if the reader is new to these notions she is encouraged to attempt her own proof before looking at the proof included at the end of this section.

We will need the following plausible but key fact arising from the completeness axiom. The reader unused to “ ϵ, δ ” arguments may find the proof a little abstract but if she draws a picture or two all will become clear.

THEOREM 2 *Every non-empty subset of the real numbers that is bounded from above has a least upper bound.*

Of course, there are companion assertions for lower bounds. If any set \mathbb{A} has a lower bound then it also has a *greatest lower bound* which is also called the *infimum* of \mathbb{A} and denoted $\inf \mathbb{A}$. If \mathbb{A} has no lower bound we write simply $\inf \mathbb{A} = -\infty$. We can obtain parallel assertions to those for the supremum by recognising that $\inf \mathbb{A} = -\sup(-\mathbb{A})$ where $-\mathbb{A}$ is the set with all the elements of \mathbb{A} negated.

The next two theorems are fundamental. I will eschew generality and provide versions sufficient for the purpose.

A *cover* of a subset \mathbb{I} of the real line is a collection $\{\mathbb{U}_n, n \geq 1\}$ of sets whose union contains \mathbb{I} ; it is an *open cover* if each \mathbb{U}_n is an open interval (a_n, b_n) . A *subcover* of a given cover is merely a subcollection whose union also contains \mathbb{I} ; it is a *finite subcover* if the subcollection contains only a finite number of sets.

HEINE–BOREL THEOREM *Suppose \mathbb{I} is a closed and bounded interval. Then every open cover $\{\mathbb{U}_n, n \geq 1\}$ of \mathbb{I} has a finite subcover.*

Now, a sequence of real numbers need have no convergence properties whatsoever. However, an important fact about sequences is that, even if a sequence is non-convergent, if it is bounded then it will always have an embedded convergent subsequence. This fact is key to the proofs of many theorems in analysis; the nearest example is in Section XXI.2.

BOLZANO–WEIERSTRASS THEOREM *Every bounded sequence has a convergent subsequence.*

Let $\{x_n, n \geq 1\}$ be any sequence of real numbers. While it is too much to expect, in general, that $\{x_n\}$ converges, we can, however, extract an upper and lower envelope for the ends of the sequence. To formalise this idea, for each n , introduce the nonce notation $\bar{x}_n = \sup\{x_n, x_{n+1}, x_{n+2}, \dots\}$. The sequence $\{\bar{x}_n, n \geq 1\}$ thus obtained is a collection of least upper bounds (possibly infinite) of the sequence $\{x_n\}$ with elements successively deleted from consideration. It follows that $\{\bar{x}_n\}$ is a decreasing sequence and consequently has a limit (possibly infinite). We call the limiting value the *limit superior* of the sequence $\{x_n\}$ and denote it by $\limsup x_n = \lim_{n \rightarrow \infty} \bar{x}_n = \inf_{n \geq 1} \sup_{m \geq n} x_m$. Likewise, if we let $\underline{x}_n = \inf\{x_n, x_{n+1}, x_{n+2}, \dots\}$ for each n , then $\{\underline{x}_n, n \geq 1\}$ is an increasing sequence of values (each possibly $-\infty$) and hence converges to a limit (possibly infinite). We call this limiting value the *limit inferior* of the sequence $\{x_n\}$ and denote it by $\liminf x_n = \lim_{n \rightarrow \infty} \underline{x}_n = \sup_{n \geq 1} \inf_{m \geq n} x_m$. No new ground is being broken here. In fact, $\liminf_n a_n = -\limsup_n (-a_n)$, as is easy to verify. Other simple properties follow as readily: if A is any number then $\liminf_n (A - a_n) = A - \limsup_n a_n$ and $\liminf_n (A + a_n) = A + \liminf_n a_n$.

As a simple illustration, the alternating-sign sequence $x_n = (-1)^n$ is non-convergent with $\limsup(-1)^n = 1$ and $\liminf(-1)^n = -1$.

It is clear that $\liminf x_n \leq \limsup x_n$ with equality if, and only if, the sequence $\{x_n\}$ is convergent. In this case all three limits coincide and $\liminf x_n = \lim x_n = \limsup x_n$. The superior and inferior limits are useful in situations where we have sequences of unknown convergence properties: if it can be established that the limits superior and inferior of the sequence at hand are equal then it must be the case that the sequence converges. The reader will find a powerful example of the use of this principle in the proof of the dominated convergence theorem in Section XIII.8.

We turn now to sequences of functions. Let $\{f_n, n \geq 1\}$ be a sequence of real-valued functions on some space Ω .

DEFINITION 3 A sequence of functions $\{f_n\}$ converges pointwise to a function f , in notation, $f_n \rightarrow f$, if, for every $\omega \in \Omega$ and every $\epsilon > 0$, there exists an integer N determined by f , ω , and ϵ , such that $|f_n(\omega) - f(\omega)| < \epsilon$ for all $n \geq N$.

In short, $f_n \rightarrow f$ if, for every ω , the sequence of values $\{f_n(\omega)\}$ converges to $f(\omega)$, or, in notation, $\lim f_n(\omega) = f(\omega)$. While a given sequence $\{f_n\}$ will not in general converge, for each $\omega \in \Omega$, the values $\sup f_n(\omega)$, $\inf f_n(\omega)$, $\limsup f_n(\omega)$, and $\liminf f_n(\omega)$ all exist (if possibly infinite) and hence determine functions $\sup f_n$, $\inf f_n$, $\limsup f_n$, and $\liminf f_n$, respectively, on the domain Ω , with the caveat that each of these functions can possibly take values $\pm\infty$ at some points of the domain. Such functions are referred to as *extended real-valued functions*. If, in particular, the sequence $\{f_n\}$ converges pointwise to a limiting function f , then $f = \liminf f_n = \lim f_n = \limsup f_n$.

Pointwise convergence is nice but it does not, however, prescribe the rate of convergence, nor does it preclude convergence at different rates at different points ω in the domain. Now wouldn't it be nice if we could arrange matters so that pointwise convergence was at the same rate, or at least that there was a bounded worst-case convergence rate, *everywhere*?

DEFINITION 4 A sequence of functions $\{f_n\}$ converges uniformly to a function f if, for every $\epsilon > 0$, there exists an integer N determined solely by f and ϵ , so that $|f_n(\omega) - f(\omega)| < \epsilon$ at all points ω whenever $n \geq N$.

What makes the convergence uniform is that the rate of convergence does not depend on the choice of ω . If we have uniform convergence then, eventually, for all sufficiently large values n , the functions f_n cosy up to f everywhere. Thus, f_n will be a uniformly good approximation to f eventually. If we have uniform convergence then, for sufficiently large n , the function f_n can be held up as a satisfactory surrogate for f .

PROOF OF THEOREM 1: Suppose first that $v = \sup \mathbb{A}$. By definition then, $v \geq x$ for all x in \mathbb{A} . Suppose now there exists $\epsilon > 0$ such that there are no points of \mathbb{A} in the interval $(v - \epsilon, v]$. Then $v - \epsilon$ is an upper bound of \mathbb{A} and we have discovered an upper bound of \mathbb{A} that is less than v which is a contradiction. Thus, for every $\epsilon > 0$ there exists x in \mathbb{A} with $x > v - \epsilon$.

Suppose on the other hand that v is an upper bound of \mathbb{A} and, for every $\epsilon > 0$, there exists $x \in \mathbb{A}$ with $x > v - \epsilon$. To establish a contradiction, suppose that $\sup \mathbb{A} = u < v$. But then $v - u > 0$. Setting $\epsilon = v - u$ we then observe that, by the given condition, there exists $x \in \mathbb{A}$ with $x > v - \epsilon = v - (v - u) = u$. But then this x is a witness for the statement that u is not an upper bound of \mathbb{A} and we have a contradiction. It follows that we must indeed have $\sup \mathbb{A} = v$. ▶

PROOF OF THEOREM 2: Suppose \mathbb{A} is a non-empty subset of real numbers bounded above. To begin, let x_0 be any point in \mathbb{A} and let v_0 be any upper bound of \mathbb{A} . Then $x_0 \leq v_0$. We now construct an increasing sequence $\{x_n\}$ of points in \mathbb{A} and a decreasing sequence $\{v_n\}$ of upper bounds of \mathbb{A} , both sequences converging to the least upper bound of \mathbb{A} . The idea here is that we recursively move points in \mathbb{A} and upper bounds of \mathbb{A} closer and closer to each other by reducing the separation between them by one-half or better at each step.

For $n \geq 0$, if $x_n = v_n$ then v_n is the least upper bound of \mathbb{A} and the sequence terminates after the n th step; we formally set $x_{n+1} = x_n = v_n = v_{n+1}$ so that neither sequence is altered from this point on. Suppose, on the other hand, $\delta_n = v_n - x_n$ is strictly positive. If there are no points of \mathbb{A} in the interval $(x_n + \delta_n/2, v_n]$, set $x_{n+1} = x_n$ and $v_{n+1} = x_n + \delta_n/2$. If, on the other hand, there are points of \mathbb{A} in the interval $(x_n + \delta_n/2, v_n]$, pick any point x_{n+1} of \mathbb{A} with $x_{n+1} > x_n + \delta_n/2$ and set $v_{n+1} = v_n$.

The construction makes clear that, in all cases, $x_n \leq x_{n+1} \leq v_{n+1} \leq v_n$. By induction it is now easy to establish that $\{x_n\}$ is an increasing sequence of points in \mathbb{A} , $\{v_n\}$ is a decreasing sequence of upper bounds of \mathbb{A} , and, for each n , $0 \leq \delta_n = v_n - x_n \leq \delta_0/2^n$. As $\{v_n\}$ is a decreasing sequence bounded below by x_0 it follows by the completeness axiom that v_n converges monotonically to some number v . We claim that v is indeed the least upper bound of \mathbb{A} . Suppose in fact that v is not an upper bound of \mathbb{A} . Then there exists x in \mathbb{A} with $x > v$ so that $\epsilon = x - v$ is

strictly positive. But as $v_n \downarrow v$, for all sufficiently large n , $v \leq v_n < v + \epsilon/2$ which means that $v_n < x = v + \epsilon$. But this is a contradiction as v_n is an upper bound of \mathbb{A} and hence $v_n \geq x$. It follows that v is indeed an upper bound of \mathbb{A} . As x_n is in \mathbb{A} , it follows that $x_n \leq v$ for each n . By the triangle inequality, it follows that, for each n , $0 \leq v - x_n \leq |v - v_n| + |v_n - x_n|$. Pick n so large that $|v - v_n| < \epsilon/2$ (possible as $v_n \downarrow v$) and $\delta_n = v_n - x_n \leq \delta_0/2^n < \epsilon/2$ (made possible by the exponential decay of $\delta_0/2^n$). Then, for all sufficiently large n , we have $0 \leq v - x_n < \epsilon$ or $x_n > v - \epsilon$. Thus, for every $\epsilon > 0$, there are points of \mathbb{A} in the interval $(v - \epsilon, v]$. It follows that v is indeed the least upper bound of \mathbb{A} . ▶

PROOF OF THE HEINE–BOREL THEOREM: Suppose $\mathbb{I} = [a, b]$ for some $a < b$ and let $\{\mathbb{U}_n, n \geq 1\}$ be any open cover of \mathbb{I} . Consider the set \mathbb{A} consisting of those real numbers $x \in [a, b]$ for which the closed interval $[a, x]$ can be covered by a finite subcollection of the \mathbb{U}_n . The set \mathbb{A} is not empty (it at least contains the point a) and is bounded below by a and above by b . It follows that $\sup \mathbb{A} = c$ lies in the interval $[a, b]$. Accordingly some element in the cover $\{\mathbb{U}_n, n \geq 1\}$, say \mathbb{U}_N , contains c . As \mathbb{U}_N is an open interval it contains within it an open interval $(c - \epsilon, c + \epsilon)$ for some suitably small positive ϵ . (If the reader draws a picture she will see what is going on.) As c is the least upper bound of the set \mathbb{A} there exists a point x in \mathbb{A} with $c - \epsilon < x < c$. It follows that the closed interval $[a, x]$ has a finite subcover, say $\mathbb{U}_1, \dots, \mathbb{U}_m$. The interval $[x, c + \epsilon/2]$ is contained in the interval $(c - \epsilon, c + \epsilon)$ and accordingly the closed interval $[a, c + \epsilon/2]$ has the finite subcover $\mathbb{U}_1, \dots, \mathbb{U}_m, \mathbb{U}_N$. We conclude that c is in \mathbb{A} and, moreover, $c = b$. Indeed, if $c < b$ then we would be able to find a member of \mathbb{A} larger than c as $[a, c + \epsilon/2]$ has a finite subcover. But this cannot happen as $c = \sup \mathbb{A}$ is the least upper bound of \mathbb{A} . ▶

PROOF OF THE BOLZANO–WEIERSTRASS THEOREM: Suppose $\{x_n, n \geq 1\}$ is a bounded sequence. We may suppose then that there is a positive M such that $|x_n| \leq M/2$ for all n . Then at least one of the two intervals $[-M/2, 0]$ and $[0, M/2]$ contains an infinite number of members of the sequence $\{x_n\}$. We may hence pick a closed subinterval of width $M/2$ which contains an infinite subsequence $\{x_{1n}, n \geq 1\}$ of the original sequence $\{x_n, n \geq 1\}$. We divide this interval into two and pick a subinterval of width $M/4$ which contains an infinite subsequence $\{x_{2n}, n \geq 1\}$ of the subsequence $\{x_{1n}, n \geq 1\}$, and, proceeding in this fashion, after k steps we will have picked a subinterval of width $M2^{-k}$ containing an infinite subsequence of points $\{x_{kn}, n \geq 1\}$. It is clear that by this process we will have created a family of nested subsequences, $\{x_{k+1,n}, n \geq 1\} \subseteq \{x_{kn}, n \geq 1\}$. Pick one member, say the first, from each of these subsequences. Then the points $x_{11}, x_{21}, \dots, x_{k1}, \dots$ form a subsequence with the property that, for each k , the set of points $\{x_{k1}, x_{k+1,1}, x_{k+2,1}, \dots\}$ is contained in the subsequence $\{x_{kn}, n \geq 1\}$. It follows that, for each k , $|x_{k1} - x_{k+m,1}| \leq M2^{-k}$ for every $m \geq 0$. As $M2^{-k}$ can be made as small as desired by choosing k sufficiently large, it follows that the members of the subsequence $\{x_{j1}, j \geq k\}$ get arbitrarily close to each other for large enough k . Or, in other words, the specified subsequence $\{x_{k1}, k \geq 1\}$ is a Cauchy sequence, hence convergent by the completeness axiom of the real numbers. ▶

2 Continuous functions

The notion of a function *qua* function is too general to do much with and we will need to put sensible “regularity” restrictions on the functions we deal with before we can make progress. The most natural of these is the idea of continuity familiar from elementary calculus. To prepare the ground we hence begin with a review of continuous functions on the real line. Orthodoxy compels us to identify points on the real line by x instead of w .

Continuity at a point conjures up the intuitive idea that the function must vary smoothly around that point or, more visually, that the graph of the function cannot

exhibit a break at that point. The historical record shows, however, that capturing this most natural of concepts mathematically proved to be very slippery. Luminaries like Newton, Leibniz, Euler, and Cauchy struggled to formally describe the concept and it was only late in the nineteenth century that Karl Weierstrass found the key and formulated the definition that we use today.

DEFINITION 1 Suppose f is a real-valued function defined on some interval \mathbb{I} , possibly infinite. Let x_0 be any point in \mathbb{I} . Then f is *continuous at x_0* if, for every $\epsilon > 0$, there exists $\delta > 0$, determined only by f , ϵ , and x_0 , such that $|f(x) - f(x_0)| < \epsilon$ whenever $x \in \mathbb{I}$ and $|x - x_0| < \delta$.

Thus, continuity at x_0 means that, if $\{x_n\}$ is any sequence of points converging to x_0 , then $f(x_n) \rightarrow f(x_0)$. In Figure 1, the function f is continuous at x_1 , continuous from the right at x_2 , and not continuous from either direction at x_3 .

A beginning student of calculus is now shown Weierstrass's definition as a matter of course and expected to master it instantaneously; but where gods struggle mortals would do well to tread cautiously—there are treacherous shoals in these waters.

In Weierstrass's definition, ϵ plays the rôle of the admissible separation of “ y ” values and δ the size of the “ x ” neighbourhood. It is critical that no limitation be placed on the size of the positive ϵ ; this ensures that no gap in the graph, however small, can squeeze through. In Figure 1, f is continuous at the point x_1 as, for any desired level of approximation ϵ , we can find a small enough neighbourhood around x_1 within which the function values differ from $f(x_1)$ by no more than ϵ . The function is discontinuous at x_2 and x_3 because *every* neighbourhood of x_2 and x_3 contains points at which the function value is “far” (that is to say, at least as large as the size of the gap away) from its value at x_2 and x_3 , respectively.

To show that a function is continuous at a given point we will have to exhibit, either explicitly or implicitly, a value δ for each choice of ϵ . The following elementary exercises will serve to illustrate the concept.

EXAMPLES: 1) *Linear forms.* The simplest, non-trivial example of a continuous function is the linear form $f(x) = ax + b$ of slope $a \neq 0$ and intercept b . For any x_0 , we then have $f(x) - f(x_0) = a(x - x_0)$ so that, for every fixed, positive ϵ , we have $|f(x) - f(x_0)| = |a| \cdot |x - x_0| < \epsilon$ whenever $|x - x_0| < \epsilon/|a| = \delta$.

2) *Quadratic functions.* If $f(x) = x^2$ then $f(x) - f(x_0) = x^2 - x_0^2 = (x - x_0)(x + x_0)$. It follows that

$$|f(x) - f(x_0)| = |x - x_0| \cdot |x + x_0| \leq |x - x_0|(|x| + |x_0|).$$

It is clear that the right-hand side can be made as small as desired if x is in a sufficiently small neighbourhood of x_0 . Formally, for $|f(x) - f(x_0)|$ to be less than ϵ it will suffice if $|x - x_0| < \epsilon/(|x| + |x_0|)$. It will suffice hence to select any neighbourhood size $0 < \delta < \min\{1, \epsilon/(2|x_0| + 1)\}$ to ensure that $|f(x) - f(x_0)| < \epsilon$.

3) *The sine and the cosine.* The reader is probably familiar with the identity $\sin x < x$ for $0 < x < \pi/2$. (If not, she should compare the area $\frac{1}{2}ab$ of the triangle OCP with

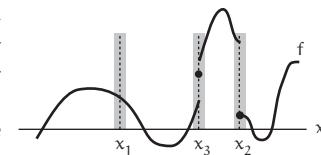


Figure 1: Illustration of continuity.

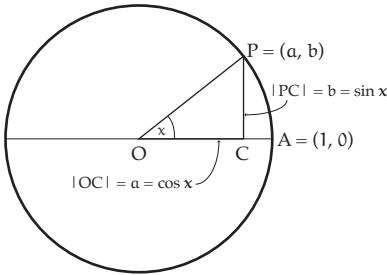


Figure 2: A proof by picture of the trigonometric inequality $\sin x < x$ for $0 < x < \pi/2$.

the area $\frac{1}{2}1^2x$ of the sector OAP in Figure 2.) The elementary trigonometric identity $\sin x - \sin x_0 = 2 \cos\left(\frac{x+x_0}{2}\right) \sin\left(\frac{x-x_0}{2}\right)$ hence shows that

$$|\sin x - \sin x_0| \leq 2|\cos\left(\frac{x+x_0}{2}\right)| \cdot |\sin\left(\frac{x-x_0}{2}\right)| \leq 2|\sin\left(\frac{x-x_0}{2}\right)| \leq |x - x_0|$$

as the cosine is bounded in absolute value by 1. It follows that for any $0 < \epsilon < \pi/2$, we have $|\sin x - \sin x_0| < \epsilon$ whenever $|x - x_0| < \epsilon$ and we may select $\delta = \epsilon$. By replacing x and x_0 by $x + \pi/2$ and $x_0 + \pi/2$, respectively, we see that similar bounds hold for the cosine. ►

Of course, we are interested in more than continuity at a single point.

DEFINITION 2 A real-valued function f on an interval \mathbb{I} is *continuous* if it is continuous at each point of \mathbb{I} .

The family of functions f that are continuous on an interval \mathbb{I} is denoted $\mathcal{C}(\mathbb{I})$ (or simply, \mathcal{C} if the interval is clear from context). The intervals of interest to us are primarily the unit interval $[0, 1]$, the real line $\mathbb{R} = (-\infty, \infty)$, and the half-line $\mathbb{R}^+ = (0, \infty)$.

It is occasionally useful to break down the idea of continuity into approaches to the point of interest from the right and from the left.

DEFINITION 3 We say that f is *continuous from the right* (or *right continuous*) at the point x_0 if $f(x_n) \rightarrow f(x_0)$ for any sequence of points $\{x_n, n \geq 1\}$ converging to x_0 from above. We capture this in notation by writing $f(x_0) = f(x_0+)$. We say that f is *right continuous* if it is continuous from the right at all points in its domain. Likewise, we say that f is *continuous from the left* (or *left continuous*) at the point x_0 if $f(x_n) \rightarrow f(x_0)$ for any sequence of points $\{x_n, n \geq 1\}$ converging to x_0 from below. We capture this in notation by writing $f(x_0) = f(x_0-)$. We say that f is *left continuous* if it is continuous from the left at all points in its domain.

It should be clear that f is continuous at x_0 if, and only if, it is both right continuous at x_0 and left continuous at x_0 . Thus, in Figure 1, f is continuous from the right at x_2 and not continuous from either side at x_3 . The simplest example of a right continuous function that is not continuous is the *Heaviside function*,

$$H_0(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

named after Oliver Heaviside who made extensive use of it in his, then radical, operator methods.

Now, intuitively, the more violently a continuous function varies around a given point x_0 , the smaller the neighbourhood (choice of δ) for which the function values stay close (the given ϵ) to the value of the function at that point. The size of neighbourhood around a point that is necessary hence depends implicitly on how the function behaves around that point. Now, wouldn't it be nice if we could find a *fixed* neighbourhood size within which we get a *uniformly good approximation* to function values at any point?

DEFINITION 4 We say that f is *uniformly continuous* if, for every $\epsilon > 0$ and x_0 in \mathbb{I} , there exists a positive δ , determined solely by f and ϵ , such that $|f(t) - f(x_0)| < \epsilon$ whenever t is in the interval \mathbb{I} and $|t - x_0| < \delta$.

What makes continuity uniform is that the positive δ may be specified independently of the chosen point x_0 .

Graphically, uniform continuity means that, for any $\epsilon > 0$, we can cut a cylinder with axis parallel to the x -axis, of diameter ϵ and width some fixed δ determined by ϵ alone, so that the cylinder can slide freely along the graph of the function without the curve hitting the sides of the cylinder. Figure 3 may help cement the mental picture.

The polynomials, sinusoids, and exponentials provide familiar elementary continuous functions on the real line while logarithms are continuous on the half-line.

Of these the graphs of the sinusoids exhibit only bounded excursions from zero while the graphs of polynomials, exponentials, and logarithms have unbounded excursions.

DEFINITION 5 A real-valued function f on an interval \mathbb{I} is *bounded* if there exists a finite value M such that $|f(t)| \leq M$ for all t in \mathbb{I} .

When does a continuous function fail to be uniformly continuous? A consideration of situations where uniformity fails may suggest what we should look for. The sinusoids $\sin x$ and $\cos x$ are bounded and uniformly continuous everywhere on the real line. The exponential e^x on the other hand is certainly continuous but not bounded or uniformly continuous as a consideration of asymptotically large values shows. It is not merely boundedness that is the issue as the function $\sin(e^x)$ is continuous and bounded trivially by 1 but is not uniformly continuous; the neighbourhoods needed for satisfactory approximation become vanishingly tiny for large values of x as the function oscillates arbitrarily fast between the values -1 and $+1$ asymptotically.

A difficulty in another direction is seen by a consideration of the function $1/x$ on the positive half-line $(0, \infty)$. While it is certainly continuous, it is neither bounded nor uniformly continuous as a consideration of the vicinity of the origin shows. Likewise, the function $\tan x$ defined on the open interval $(-\pi/2, \pi/2)$ is again continuous but not bounded or uniformly continuous as its explosive growth at the endpoints shows. Still another example is provided by the function $\sin(x^{-2})$ defined on the positive half-line $(0, \infty)$ which is bounded and continuous but not uniformly continuous as the oscillations near zero show.

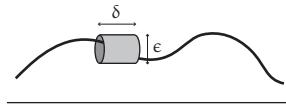


Figure 3: A uniformly continuous function.

In the settings outlined above, the slope of the offending function gets unboundedly large—near infinity in the one instance and near the endpoint of an open interval in the other. This suggests that placing limitations on the slope of a continuous function will suffice to ensure uniform continuity. Roughly speaking, a function that is uniformly continuous cannot change or wiggle “infinitely fast”. The following result is representative and the reader may wish to try her hand at constructing a proof before looking at the one I’ve provided at the end of this section.

THEOREM 1 *Suppose f is continuous on an interval \mathbb{I} and has a bounded derivative everywhere in the interior of \mathbb{I} . Then f is uniformly continuous on \mathbb{I} .*

The reader may object that possessing a bounded derivative places a heavy smoothness requirement on a function and it seems almost unfair to prove a uniform smoothness property by tacking on even more smoothness. And it is certainly true that there is a very large subclass of continuous functions that are not differentiable. A re-examination of the above examples to see whether there is anything structural in the domain that causes unbounded behaviour provides two suggestive nuggets: (i) infinite domains can pose problems; (ii) problems can arise at the open endpoints of intervals. This suggests that the problems with unboundedness may vanish if we restrict attention to closed and bounded intervals. And indeed they do.

THEOREM 2 *If f is continuous on a closed and bounded interval $[a, b]$ then it is uniformly continuous on $[a, b]$ and, moreover, achieves its maximum and minimum values on $[a, b]$.*

It follows that the restriction of a continuous function on the real line to a closed and bounded interval is uniformly continuous and bounded on that interval. The reader who is willing to take the result at face value can skip blithely on. For the insatiably curious the proof is included at the end of the section.

The idea of uniform continuity may be extended to characterise the simultaneous smoothness of entire families of functions. Suppose Λ is any index set (countable or uncountable) and \mathbb{I} any interval.

DEFINITION 6 A family $\{f_\lambda, \lambda \in \Lambda\}$ of real-valued functions defined on \mathbb{I} is *equicontinuous* if each f_λ is uniformly continuous on \mathbb{I} and, moreover, for each $\epsilon > 0$, $x_0 \in \mathbb{I}$, and $\lambda \in \Lambda$, we may select a uniform, positive δ , determined only by ϵ and the chosen family of functions, and independent of x_0 and λ , such that $|f_\lambda(x) - f_\lambda(x_0)| < \epsilon$ whenever $x \in \mathbb{I}$ and $|x - x_0| < \delta$.

We may express uniform continuity and equicontinuity vividly in terms of the notion of the oscillation of the function. If \mathbb{I} is any interval, the *oscillation of f over \mathbb{I}* is defined by $\text{osc}_{\mathbb{I}} f := \sup\{|f(x) - f(y)| : x \in \mathbb{I}, y \in \mathbb{I}\}$. Then, a function f is uniformly continuous if, for some $\delta = \delta(\epsilon; u)$, the oscillation of f is $< \epsilon$ over every interval of width δ . And, likewise, a family $\{f_\lambda\}$ is equicontinuous if, for some $\delta = \delta(\epsilon; \Lambda)$, the oscillation of f_λ over any interval of width δ is less than ϵ for every member of the family.

We now turn to a consideration of sequences of functions. It should not be too surprising that the uniform convergence of a sequence of uniformly continuous functions preserves smoothness.

THEOREM 3 Suppose $\{f_n\}$ is a sequence of continuous functions converging uniformly to a function f on a closed and bounded interval $[a, b]$. Then f is continuous.

COROLLARY 1 Under the conditions of Theorem 3, $\int_a^b f_n(x) dx \rightarrow \int_a^b f(x) dx$ as $n \rightarrow \infty$.

COROLLARY 2 Suppose $\{f_n\}$ is a sequence of continuously differentiable functions. Suppose $f_n \rightarrow f$ pointwise and $f'_n \rightarrow g$ uniformly on each closed and bounded interval $[a, b]$. Then f is differentiable with continuous derivative g .

The notions of continuity and convergence carry over to real-valued (or even vector-valued) functions of two or more dimensions with the natural replacement of the idea of distance $|x - y|$ between points on the line by any metric in Euclidean space \mathbb{R}^n , the one most commonly in use being the Euclidean distance $\|x - y\| = \sqrt{\sum_{j=1}^n (x_j - y_j)^2}$ between points x and y in \mathbb{R}^n . Naturally enough, in this setting, intervals \mathbb{I} on the line are replaced by n -dimensional rectangles $\mathbb{I}_1 \times \cdots \times \mathbb{I}_n$ in \mathbb{R}^n obtained as the Cartesian product of n intervals.

To conclude, the reader is probably aware that a cavalier interchange of integration and differentiation operations without a glance at attendant conditions can be treacherous. Difficulties can be smoothed over if there is uniform convergence of the integral tails. A result of this stripe in the differentiation of convolution integrals is provided in the following theorem.

THEOREM 4 Suppose u is continuous and bounded and f is continuously differentiable with both f and f' integrable. Suppose that $\int_{|x| \geq n} |f'(t-x)| dx \rightarrow 0$ as $n \rightarrow \infty$ uniformly in t on every closed and bounded interval $[a, b]$. Then $(f * u)' = f' * u$. In detail,

$$\frac{d}{dt} \int_{-\infty}^{\infty} u(x)f(t-x) dx = \int_{-\infty}^{\infty} u(x)f'(t-x) dx.$$

PROOF OF THEOREM 1: Pick any point x_0 in \mathbb{I} . The mean value theorem of calculus then says that, for any point x in \mathbb{I} there is at least one interior point c between x_0 and x for which $f(x) - f(x_0) = f'(c)(x - x_0)$. [In words: there is an interior point c at which the slope of the function matches the slope of the chord connecting the points $(x_0, f(x_0))$ and $(x, f(x))$.] As f' is bounded in absolute value, say by M , *a fortiori* we have $|f'(c)| < M < \infty$. Consequently, $|f(x) - f(x_0)| < M|x - x_0|$. It follows that $|f(x) - f(x_0)| < \epsilon$ whenever $|x - x_0| < \epsilon/M$ and we may select $\delta = \epsilon/M$ independently of the point x_0 in the definition of uniform continuity. ▶

PROOF OF THEOREM 2: Suppose that $f: [a, b] \rightarrow \mathbb{R}$ is continuous but not uniformly continuous on $[a, b]$. Then for some $\epsilon > 0$ and any $\delta > 0$ there exist points x and y in $[a, b]$ satisfying $|x - y| < \delta$ and $|f(x) - f(y)| > \epsilon$. By successively choosing as δ the values of the sequence $\{1/n, n \geq 1\}$, it follows that there exists a sequence of pairs $\{(x_n, y_n), n \geq 1\}$ with $|x_n - y_n| < 1/n$ and $|f(x_n) - f(y_n)| > \epsilon$. As $\{x_n\}$ is bounded it follows by the Bolzano–Weierstrass theorem that it has a convergent subsequence $\{x_{n_i}, i \geq 1\}$ with x_{n_i} converging to some point x_0 in the interval $[a, b]$. Now consider the subsequence $\{y_{n_i}, i \geq 1\}$. As it is bounded it has a convergent subsequence $\{y_{n_{i_j}}, j \geq 1\}$ with $y_{n_{i_j}}$ converging to some point y_0 in the interval $[a, b]$. But as $\{x_{n_{i_j}}, j \geq 1\}$ is a subsequence of the convergent sequence $\{x_{n_i}, i \geq 1\}$, it must converge to the same limit and so $x_{n_{i_j}} \rightarrow x_0$ as $j \rightarrow \infty$. We may express y_0 by means of the telescoping sum

$$y_0 = x_0 + (x_{n_{i_j}} - x_0) + (y_{n_{i_j}} - x_{n_{i_j}}) + (y_0 - y_{n_{i_j}})$$

and as each of the three terms in round brackets on the right tends to zero as j tends to infinity, we are left with the conclusion that $y_0 = x_0$ identically.

Now the continuity of the function f means that $f(x_{n_i}) \rightarrow f(x_0)$ and $f(y_{n_i}) \rightarrow f(y_0)$. Two applications of the triangle inequality show then that

$$\begin{aligned} |f(x_{n_i}) - f(y_{n_i})| &\leq |f(x_{n_i}) - f(x_0)| + |f(x_0) - f(y_0)| + |f(y_0) - f(y_{n_i})| \\ &= |f(x_{n_i}) - f(x_0)| + |f(y_0) - f(y_{n_i})| < \epsilon \end{aligned}$$

eventually, contradicting the original hypothesis. It follows that f must be uniformly continuous on $[a, b]$.

To finish the proof, let $S = \sup\{f(x) : a \leq x \leq b\}$. Then there is a sequence $\{x_n\}$ of points in $[a, b]$ such that $f(x_n) \rightarrow S$. The sequence $\{x_n\}$ may not be convergent but, as it is bounded, it must have a convergent subsequence $\{x_{n_i}\}$ with x_{n_i} converging to some point x_0 in $[a, b]$. As f is continuous, it follows that $f(x_{n_i}) \rightarrow f(x_0)$. On the other hand, $\{f(x_{n_i})\}$ is a subsequence of the convergent sequence $\{f(x_n)\}$ and so it must converge to the same limit S . It follows that $f(x_0) = S$ and f indeed achieves its maximum value in the interval $[a, b]$. Arguing as in the above with f replaced by $-f$ will establish that f achieves its minimum value as well. ▶

PROOF OF THEOREM 3: As the sequence f_n converges uniformly, we may select a sufficiently large n so that $|f_n(x) - f(x)| < \epsilon$ for all $x \in [a, b]$. As f_n is continuous, we may select $\delta > 0$ so that $|f_n(x) - f_n(y)| < \epsilon$ for all $y \in [a, b]$ with $|x - y| < \delta$. It follows by the triangle inequality that

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)| < 3\epsilon$$

whenever $|x - y| < \delta$. As $\epsilon > 0$ may be taken arbitrarily small, it follows that f is continuous and consequently also uniformly continuous on $[a, b]$. ▶

PROOF OF COROLLARY 1: As f is uniformly continuous on $[a, b]$, it follows that

$$\left| \int_a^b f_n(x) dx - \int_a^b f(x) dx \right| \leq \int_a^b |f_n(x) - f(x)| dx < (b-a)\epsilon,$$

eventually. As ϵ may be chosen arbitrarily small, the desired conclusion follows. ▶

PROOF OF COROLLARY 2: In view of Corollary 1, we have

$$\int_a^x f'_n(\xi) d\xi \rightarrow \int_a^x g(\xi) d\xi \quad (n \rightarrow \infty).$$

On the other hand, as the sequence $\{f_n\}$ converges pointwise, we have

$$\int_a^x f'_n(\xi) d\xi = f_n(x) - f_n(a) \rightarrow f(x) - f(a) \quad (n \rightarrow \infty).$$

The limits on the right must coincide and so

$$f(x) - f(a) = \int_a^x g(\xi) d\xi.$$

The function g is continuous by Theorem 3 and so, by taking derivatives of both sides, the fundamental theorem of calculus shows that f is differentiable and $f'(x) = g(x)$. ▶

PROOF OF THEOREM 4: Introduce the nonce notation

$$h_n(t) = \int_{-n}^n u(x)f(t-x) dx, \quad h(t) = \int_{-\infty}^{\infty} u(x)f(t-x) dx.$$

We may suppose that $|u| \leq M$ for some M and so

$$\begin{aligned} |h_n(t) - h(t)| &= \left| \int_{|x| \geq n} u(x)f(t-x) dx \right| \\ &\leq \int_{|x| \geq n} |u(x)| |f(t-x)| dx \leq M \int_{|x| \geq n} |f(t-x)| dx \rightarrow 0 \end{aligned}$$

as the integrability of f implies that the area in the tails must tend to zero. It follows that $h_n \rightarrow h$ pointwise. We now consider the function

$$H_n(t) = \int_{-n}^n u(x)f'(t-x) dx.$$

As f' is uniformly continuous on the closed and bounded set $[-n, n]$ and u is bounded absolutely by M , it is easy to see that $H_n(t)$ is continuous. Indeed, for any $\epsilon > 0$ we may select $\delta > 0$ so that $|f'(t) - f'(s)| < \epsilon$ for all $-n \leq t, s \leq n$ with $|t - s| < \delta$. Accordingly,

$$\begin{aligned} |H_n(t) - H_n(s)| &= \left| \int_{-n}^n u(x)(f'(t-x) - f'(s-x)) dx \right| \\ &\leq M \int_{-n}^n |f'(t-x) - f'(s-x)| dx \leq 2Mn\epsilon, \end{aligned}$$

and as ϵ may be chosen arbitrarily small, it follows that H_n is continuous. The integrands are all well-behaved (uniformly continuous and bounded in every closed and bounded interval) and accordingly we may swap the order of integration to obtain

$$\int_0^t H_n(\tau) d\tau = \int_{-n}^n u(x) \int_0^t f'(\tau-x) d\tau dx = \int_{-n}^n u(x)(f(t-x) - f(-x)) dx = h_n(t) - h_n(0).$$

As H_n is continuous, we may differentiate both sides whence the fundamental theorem of calculus tells us that h_n is continuously differentiable with $h'_n(t) = H_n(t)$. To mop up, we observe that

$$\left| h'_n(t) - \int_{-\infty}^{\infty} u(x)f'(t-x) dx \right| = \left| \int_{|x| \geq n} u(x)f'(t-x) dx \right| \leq M \int_{|x| \geq n} |f'(t-x)| dx \rightarrow 0$$

uniformly on every closed and bounded interval. It follows by Corollary 2 that $h(t) = (f * u)(t)$ is differentiable with $(f * u)'(t) = (f' * u)(t)$. ▶

3 Some L^2 function theory

We conclude this Appendix with a brief excursus on complete orthonormal sequences. We will deal with the family of complex-valued, square-integrable functions defined on the unit interval $[0, 1]$. These are the complex-valued functions f for which the integral $\int_0^1 |f(x)|^2 dx$ converges. The family of all such functions is given the fancy name L^2 (or $L^2[0, 1]$ if it is desirable to make the domain of definition explicit). The family L^2 is sufficiently rich for our purposes and contains, in particular, the family of bounded continuous functions on the unit interval $[0, 1]$.

Functions in the class L^2 can be added and scaled by complex scalars and, in this respect, behave like ordinary Euclidean vectors as shown in the list of correspondences given in Table 1. These analogies suggest a geometric perspective where each function f may be thought of as an “arrow” or “vector” embedded in the abstract space of functions L^2 with an orientation specified with respect to the function that takes value identically zero which serves as the origin of the space. This geometric intuition, while pretty, will become much more potent if other attributes familiar from Euclidean geometry such as coordinate axes, vector lengths, projections of one vector onto another, the angle between two vectors, the Pythagorean theorem, and the triangle inequality have proper analogues in this mysterious new space. We begin by considering projections; if z is complex then, as usual, z^* denotes its complex conjugate.

	Euclidean Vectors		Functions	
	Object	Representation	Object	Representation
Vector	x	(x_1, x_2, x_3)	f	$f(x)$
Zero	0	$(0, 0, 0)$	0	$f(x) = 0$
Sum	$x + y$	$(x_1 + y_1, x_2 + y_2, x_3 + y_3)$	$f + g$	$f(x) + g(x)$
Scale	λx	$(\lambda x_1, \lambda x_2, \lambda x_3)$	λf	$\lambda f(x)$

Table 1: Correspondences between Euclidean and function spaces.

DEFINITION 1 Suppose f and g are square-integrable functions on the unit interval $[0, 1]$. The *inner product* between f and g is defined by

$$\langle f, g \rangle := \int_0^1 f(x)g(x)^* dx \quad (3.1)$$

and defines a map from an ordered pair of functions into the complex plane.

There is certainly a visceral similarity between this definition and the usual idea of a vector projection as a sum of componentwise products. Do the basic properties of projection hold? Yes, indeed. The reader will be readily able to verify that the inner product satisfies the natural properties we associate with projections. (1) *Conjugate symmetry*: $\langle g, f \rangle = \langle f, g \rangle^*$. (2) *Additivity*: $\langle f + g, h \rangle = \langle f, h \rangle + \langle g, h \rangle$. (3) *Homogeneity*: $\langle \lambda f, g \rangle = \lambda \langle f, g \rangle$. (4) *Positivity*: $\langle f, f \rangle \geq 0$ with equality if, and only if, $f = 0$ a.e. (see Definition V.7.2). To obviate a needless terminological nuisance we agree henceforth to identify two functions that agree almost everywhere.

The essential nature of inner products is captured in the following famous inequality which the reader will probably recognise from her physics classes.

THE CAUCHY–SCHWARZ INEQUALITY Suppose f and g are square-integrable functions on the unit interval $[0, 1]$. Then $|\langle f, g \rangle| \leq \sqrt{\langle f, f \rangle} \cdot \sqrt{\langle g, g \rangle}$. Equality can hold if, and only if, $\lambda f + g = 0$ for some complex scalar λ .

The next step in cementing our geometric intuition about projections in this setting is to articulate a notion of length of functions.

DEFINITION 2 Suppose f is any square-integrable function on the unit interval. The *length* or *norm* of f is defined by $\|f\| := \sqrt{\langle f, f \rangle}$ and specifies a map from the space L^2 of square-integrable functions into the positive reals.

The reader trained on physical intuition may observe that the square of the length of a function may be identified with the physical concept of its *energy* $\|f\|^2 = \int_0^1 |f(x)|^2 dx$ that plays a key rôle in physics and engineering. In this notation the Cauchy–Schwarz inequality takes on the striking, simple form $|\langle f, g \rangle| \leq \|f\| \cdot \|g\|$.

We feel instinctively that any satisfactory definition of length must gel with intuition built upon Euclidean spaces. (1) *Positivity*: length is always positive, $\|f\| \geq 0$ with equality if, and only if, $f = 0$. (2) *Scaling law*: arrows in the same “direction” have

proportional lengths, $\|\lambda f\| = |\lambda| \cdot \|f\|$. (3) *Triangle inequality*: the sum of two sides of a triangle is greater than the third, $\|f + g\| \leq \|f\| + \|g\|$ with equality if, and only if, $\lambda f + g = 0$ for some λ . The only non-trivial assertion is the important triangle inequality and this is an immediate consequence of the Cauchy-Schwarz inequality.

The geometric notions are coming together very prettily. The validity of the triangle inequality suggests that all our usual intuition about triangles continues to be well-founded in the space of square-integrable functions L^2 .

Paralleling ideas for sequences of numbers, we may introduce the idea of convergence of sequences of functions with respect to the L^2 -norm.

DEFINITION 3 Suppose $\{f_n, n \geq 1\}$ is a sequence of functions in L^2 . We say that $\{f_n\}$ converges to f in L^2 (or, in *mean-square*) if there exists $f \in L^2$ such that $\|f_n - f\| \rightarrow 0$.

Working by analogy, we may also define the Cauchy property for sequences in L^2 .

DEFINITION 4 We say that the sequence $\{f_n\}$ is *Cauchy in L^2* if, for every $\epsilon > 0$, there exists $N = N(\epsilon)$ such that, if $m > N$ and $n > N$, then $\|f_m - f_n\| < \epsilon$.

A key fact about the space L^2 is that it is *complete* in the following very natural sense.

THEOREM 2 Every Cauchy sequence in L^2 converges in mean-square to an element of L^2 .

The proof of this theorem contains, as we shall see, a particular observation that is of independent interest. For the definition of a property that holds a.e. (almost everywhere) see Section XI.3.

THEOREM 3 Every Cauchy sequence in L^2 contains a subsequence that converges a.e. to a point in L^2 .

Theorems 2 and 3 are true more generally in L^p spaces but we shall resist the temptation to investigate further.

Orthogonality plays a particularly important rôle in the picture. We again work by analogy.

DEFINITION 5 Functions f and g in L^2 are *orthogonal* if $\langle f, g \rangle = 0$.

The concept of orthogonality immediately brings to mind right triangles and the theorem of Pythagoras. Let us check it out. Suppose f and g are orthogonal. Then, if we are lucky, the functions f , g , and $f + g$ should form a right triangle with f and g as sides and $f + g$ serving as hypotenuse. This is a pretty picture but is it valid? To check the Pythagorean theorem in this setting, all we need is linearity of inner product:

$$\|f + g\|^2 = \langle f + g, f + g \rangle = \langle f, f \rangle + \langle f, g \rangle + \langle g, f \rangle + \langle g, g \rangle = \|f\|^2 + \|g\|^2.$$

Thus, the space of square-integrable functions equipped with an inner product and a norm behaves in many respects just like ordinary geometric Euclidean space. If it walks like a duck and quacks like a duck, it must be a duck. We will hence be able to freely parlay geometric intuition and think of these functions as vectors or arrows.

If f is any non-zero, square-integrable function, then its scaled version $f/\|f\|$ is a function of unit norm in the “direction” of f . It is natural to hence interpret the unit vector $f/\|f\|$ as establishing a coordinate direction. Orthogonality allows us to define “coordinate axes” for the function space L^2 .

DEFINITION 6 A family of square-integrable functions $\{\varphi_n\}$ forms an *orthonormal system* if $\|\varphi_n\| = 1$ for each n and $\langle \varphi_m, \varphi_n \rangle = 0$ whenever $m \neq n$.

In other words, an orthonormal system $\{\varphi_n\}$ is a family of unit-length (hence unit-energy) functions which are mutually orthogonal.

EXAMPLES: 1) *The Fourier basis.* The exponential system $\{e^{i2\pi nx}, n \in \mathbb{Z}\}$.

2) *The Rademacher functions.* The Rademacher system $\{r_n(x), n \geq 1\}$ of Section V.2.

3) *The Haar wavelets.* The Haar system $\{h_n(x), n \geq 0\}$ of Section X.10. ▶

Suppose that $\{\varphi_n, n \geq 0\}$ is an orthonormal system of functions. For any square-integrable f it is now natural to interpret $\hat{f}_n = \langle f, \varphi_n \rangle$ as the projection of f along the n th coordinate direction, an interpretation that is reinforced by the following theorem. For each $N \geq 0$, let the function $S_N(f, \cdot) = \sum_{n=0}^N \hat{f}_n \varphi_n$ represent the projection of f into the subspace spanned by $\varphi_0, \varphi_1, \dots, \varphi_N$.

THE MEAN-SQUARE APPROXIMATION THEOREM Suppose $\{\varphi_n, n \geq 0\}$ is an orthonormal system and f is square-integrable. For each N , let $f_N = \sum_{n=0}^N \lambda_n \varphi_n$ be in the linear span of $\varphi_0, \varphi_1, \dots, \varphi_N$. Then the mean-square approximation error $\|f - f_N\|$ is minimised if, and only if, $f_N = S_N(f, \cdot)$, that is to say, $\lambda_n = \hat{f}_n$ for each n .

The question now arises whether the sequence of functions $S_N(f, \cdot)$ forms a satisfactory representation of f if N is sufficiently large, that is to say, whether $\|f - S_N(f, \cdot)\| \rightarrow 0$ as $N \rightarrow \infty$. If this is the case then, by selecting $f_N = S_N(f, \cdot)$, we have $\|f - S_N(f, \cdot)\|^2 = \|f\|^2 - \sum_{n=0}^N \hat{f}_n^2$ whence, by letting N tend to infinity, we obtain $\|f\|^2 = \sum_{n=0}^{\infty} \hat{f}_n^2$. In words, energy is conserved in the projections of the function along the orthogonal coordinates determined by the given orthonormal system. The following famous equation codifies this concept in slightly more generality.

PARSEVAL'S EQUATION Suppose $\{\varphi_n, n \geq 0\}$ is an orthonormal system. For any $f, g \in L^2$ let $\hat{f}_n = \langle f, \varphi_n \rangle$ and $\hat{g}_n = \langle g, \varphi_n \rangle$ for each n . If $S_N(f, \cdot)$ and $S_N(g, \cdot)$ converge in mean-square to f and g , respectively, then $\langle f, g \rangle = \sum_{n=0}^{\infty} \hat{f}_n \hat{g}_n^*$ [and a fortiori $\|f\|^2 = \sum_{n=0}^{\infty} |\hat{f}_n|^2$].

In view of the mean-square approximation theorem, Parseval's equation will hold if there exist real sequences $\{\mu_n, n \geq 0\}$ and $\{\lambda_n, n \geq 0\}$ such that the sequences of functions $f_N = \sum_{n \leq N} \mu_n \varphi_n$ and $g_N = \sum_{n \leq N} \lambda_n \varphi_n$ converge in mean-square to f and g , respectively, that is, $\|f - f_N\| \rightarrow 0$ and $\|g - g_N\| \rightarrow 0$ as $N \rightarrow \infty$.

For a given orthonormal system $\{\varphi_n, n \geq 0\}$, the validity of Parseval's equation depends upon the mean-square convergence of the sequences of partial sums $S_N(f, \cdot)$ and $S_N(g, \cdot)$. These investigations are quite involved in general and account for the delicacy of the calculations needed to show that the trigonometric system is complete. The Haar system explored in Section X.10 shows how function localisation can make a huge difference in these calculations.

PROOF OF THE CAUCHY–SCHWARZ INEQUALITY: The proof of Hölder's inequality in Section XVII.3 leverages convexity and provides one of many proofs of the Cauchy–Schwarz inequality. Here is another charming classical proof due to Schwarz which relies upon the idea of completing squares.²

We may as well suppose $f \neq 0$ as the inequality is trivial if f or g is zero. Then $\langle f, f \rangle$ is strictly positive; let $\sqrt{\langle f, f \rangle}$ be the positive square root. Now let λ be any complex value. Schwarz's clever idea was to consider the function $\lambda f + g$ with the introduction of λ giving us some elbow room—it plays the rôle of a free parameter which we can later optimise. We have $0 \leq \langle \lambda f + g, \lambda f + g \rangle = |\lambda|^2 \langle f, f \rangle + \lambda^* \langle f, g \rangle + \lambda \langle f, g \rangle^* + \langle g, g \rangle$ by virtue of positivity, conjugate symmetry, and linearity of inner product. It is irresistible to complete the inviting square on the right and we hence obtain

$$0 \leq \langle \lambda f + g, \lambda f + g \rangle = \left| \lambda \sqrt{\langle f, f \rangle} + \frac{\langle f, g \rangle}{\sqrt{\langle f, f \rangle}} \right|^2 - \frac{|\langle f, g \rangle|^2}{\langle f, f \rangle} + \langle g, g \rangle.$$

The dependence on the slack variable λ is now isolated and we are free to optimise it. It is clear that the first term on the right is positive so that its smallest possible value is zero—which value it takes for the clever choice $\lambda = -\langle f, g \rangle / \langle f, f \rangle$. With this selection for λ , the first term on the right vanishes leaving us with $0 \leq \langle g, g \rangle - |\langle f, g \rangle|^2 / \langle f, f \rangle$ which is the result to be shown. Finally, for equality to hold, we must have $0 = \langle \lambda f + g, \lambda f + g \rangle$, but by the uniqueness of zero property this can only happen if $\lambda f + g = 0$. ▶

The proof of the completeness of the space L^2 requires Fatou's lemma from Section XIII.8. There is no danger of a circular argument as Fatou's lemma rests upon the monotone convergence theorem, the proof of which is independent of L^2 theory.

PROOF OF THEOREM 2: Suppose $\{f_n, n \geq 1\}$ is a Cauchy sequence in L^2 . There then exists a subsequence of integers n_1, n_2, \dots such that $\|f_{n_{j+1}} - f_{n_j}\| < 2^{-j}$ for each $j \geq 1$. The subsequence $\{f_{n_j}, j \geq 1\}$ now suggests a possible candidate function as limit. We may write $f_{n_k} = f_{n_1} + \sum_{j=1}^{k-1} (f_{n_{j+1}} - f_{n_j})$ as the series on the right telescopes. We are accordingly led to study the behaviour of the series $f = f_{n_1} + \sum_{j=1}^{\infty} (f_{n_{j+1}} - f_{n_j})$.

Set $g_k = \sum_{j=1}^k |f_{n_{j+1}} - f_{n_j}|$. By repeated applications of the triangle inequality, $\|g_k\| \leq \sum_{j=1}^k \|f_{n_{j+1}} - f_{n_j}\| < \sum_{j=1}^k 2^{-j} < 1$. By Fatou's lemma, it follows that

$$\int_0^1 \liminf g_k(x)^2 dx \leq \liminf \int_0^1 g_k(x)^2 dx = \liminf \|g_k\|^2 \leq 1.$$

But $\{g_k\}$ is an increasing sequence of positive functions and so writing $g = \liminf g_k = \lim g_k = \sum_{j=1}^{\infty} |f_{n_{j+1}} - f_{n_j}|$ we see that $\|g\| \leq 1$. In particular, this implies that g is finite a.e. [Assume not. Then the set of points x in the unit interval on which $g(x) = \infty$ constitutes a measurable subset A of strictly positive (Lebesgue) measure, $\lambda(A) > 0$. For each positive M , let A_M be the subset of points x in the unit interval on which $g(x) \geq M$. The sequence $\{A_M, M \geq 1\}$ is decreasing with limit $\bigcap_M A_M = A$, whence $\lambda(A_M) \geq \lambda(A) > 0$. Now $g \geq g1_{A_M}$ so that $\|g\|^2 \geq \|g1_{A_M}\|^2 = \int_{A_M} g(x)^2 dx \geq M\lambda(A_M) \geq M\lambda(A)$. As M can be taken arbitrarily large and $\lambda(A)$ is strictly positive this implies that $\|g\| = \infty$. Contradiction.]

Thus, the series $\sum_{j=1}^{\infty} (f_{n_{j+1}} - f_{n_j})$ converges absolutely a.e. and, in consequence, as $j \rightarrow \infty$, it follows that $f_{n_j} \rightarrow f$ a.e. To complete the specification, set $f(x) = 0$ on the measure-zero set where the series fails to converge. We now have a bona fide candidate function and it only remains to verify that $\{f_n\}$ converges in mean-square to it.

Fix any $\epsilon > 0$. As the limit inferior and the limit coincide when the latter exists, by another application of Fatou's lemma, we have

$$\int_0^1 (f_n(x) - f(x))^2 dx = \int_0^1 \liminf_j (f_n(x) - f_{n_j}(x))^2 dx \leq \liminf_j \int_0^1 (f_n(x) - f_{n_j}(x))^2 dx < \epsilon^2$$

²H. A. Schwarz, "Ueber ein die Flächen kleinsten Flächeninhalts betreffendes Problem der Variationsrechnung", *Acta Societatis Scientiarum Fennicæ*, vol. XV, pp. 318–362, 1885.

for all sufficiently large n as $\{f_n\}$ is a Cauchy sequence in L^2 . So $\|f_n - f\| < \epsilon$, eventually, and hence $\|f_n - f\| \rightarrow 0$. In particular, $f_n - f$ belongs to L^2 and, as we may write $f = f_n + (f - f_n)$, it follows that the limit function f is also square-integrable. ►

PROOF OF THEOREM 3: By construction, the subsequence $\{f_{n_j}, j \geq 1\}$ in the proof of Theorem 2 converges a.e. to f in L^2 . ►

PROOF OF THE MEAN-SQUARE APPROXIMATION THEOREM: By leveraging linearity of the inner product and collecting terms we obtain

$$\begin{aligned} \|f - f_N\|^2 &= \|f\|^2 - \sum_{n=0}^N \lambda_n^* \langle f, \varphi_n \rangle - \sum_{n=0}^N \lambda_n \langle \varphi_n, f \rangle + \sum_{m=0}^N \sum_{n=0}^N \lambda_m \lambda_n^* \langle \varphi_m, \varphi_n \rangle \\ &= \|f\|^2 + \sum_{n=0}^N [|\lambda_n|^2 - \lambda_n^* \hat{f}_n - \lambda_n \hat{f}_n^*] = \|f\|^2 + \sum_{n=0}^N |\lambda_n - \hat{f}_n|^2 - \sum_{n=0}^N |\hat{f}_n|^2, \end{aligned}$$

the final step following by completing squares. The sum of squares in the middle term on the right can only be positive and it follows that $\|f - f_N\|^2 \geq \|f\|^2 - \sum_{n=0}^N |\hat{f}_n|^2$, equality obtained only when $\lambda_n = \hat{f}_n$ for each n . ►

PROOF OF PARSEVAL'S EQUATION: Exploiting linearity of inner product we can write

$$\begin{aligned} \langle f, g \rangle &= \langle f - S_N(f, \cdot), g \rangle - \langle f - S_N(f, \cdot), g - S_N(g, \cdot) \rangle \\ &\quad + \langle f, g - S_N(g, \cdot) \rangle + \langle S_N(f, \cdot), S_N(g, \cdot) \rangle. \end{aligned} \quad (3.2)$$

Three applications of the Cauchy–Schwarz inequality now show that

$$\begin{aligned} |\langle f - S_N(f, \cdot), g \rangle| &\leq \|f - S_N(f, \cdot)\| \cdot \|g\| \rightarrow 0, \\ |\langle f - S_N(f, \cdot), g - S_N(g, \cdot) \rangle| &\leq \|f - S_N(f, \cdot)\| \cdot \|g - S_N(g, \cdot)\| \rightarrow 0, \\ |\langle f, g - S_N(g, \cdot) \rangle| &\leq \|f\| \cdot \|g - S_N(g, \cdot)\| \rightarrow 0, \end{aligned}$$

and the first three terms on the right of (3.2) hence tend to zero. For the last term we may appeal once more to orthogonality to obtain

$$\langle S_N(f, \cdot), S_N(g, \cdot) \rangle = \sum_{m=0}^N \sum_{n=0}^N \hat{f}_m \hat{g}_n \langle \varphi_m, \varphi_n \rangle = \sum_{n=0}^N \hat{f}_n \hat{g}_n.$$

Allowing N to tend to infinity on both sides of (3.2) completes the proof. ►

Index

- 1-trick (use with Cauchy-Schwarz inequality), 520
- A posteriori* probability, 65, 66
- A priori* probability, 65, 66
- a.e. (a.c., a.s.), *see* Almost everywhere
- Absolutely continuous distribution, 403, 508, 558
- Accumulated entrance fee, 585, 587, 605
- Accumulation point, 712
- Additivity
- of expectation, 216, 222, 234, 326, 328, 333, 334, 342, 429, 431, 433, 446–449, 451–455, 458, 460, 462, 475, 477, 502, 503, 517, 552, 569, 583, 589, 661, 662, 703, 722, 733, 742, 746, 752, 756, 757
- of Laplace transforms, 531
- of measure, 16, 18, 19, 22, 23, 37, 41, 42, 47, 48, 56, 72, 89, 93, 97, 202, 203, 256, 267, 289, 302, 372, 373, 378, 390, 398, 552, 657, 714
- of variance, 216, 234, 477, 743, 746
- Adjacent vertices in a graph, 102, 394
- Admissible
- function, 732, 733
 - permutation, 133
- Affine shift, 716
- Aldous, D., 651n
- Algebra of events, 14–16, 33, 384
- Alm, S. E., 686n
- Almost everywhere, 156, 380, 407, 441, 448, 460, 462, 464, 593, 595, 596, 599, 600, 784, 785, 787
- Aloha protocol, 276, 277, 761
- Altitude, 518
- Ancestor (of vertex), 687
- Ancillary
- permutation, 676
 - random variables, 597, 670, 673
- André, D., 246
- Anomalous numbers, 290
- Antheil, G., 138n
- Approximation
- by simple functions, 439, 442–444, 463, 466
 - by smooth functions, 698
 - of measure, 392
 - of sum by integrals, 60, 69, 204, 458, 591, 649, 742, 746
- Arc length distribution, 281
- Arc sine
- density, 206, 307
 - distribution, 307, 690
 - law, 205, 250, 307, 555, 690
- Arithmetic
- distribution, 198, 197–200, 508, 557, 710
 - mean, 481, 499, 500, 505
 - random variable, 197, 401, 449, 654, 656
 - sequence, 112, 113, 130, 131
- Arzelà, C., 707n
- Arzelà–Ascoli theorem, 707, 708, 716
- Ascoli, G., 707n
- Associative memory, 193
- Asymptotic expansion, 363
- Atoms of a discrete distribution, 401
- Authalic problem, 497
- Auxiliary randomisation, 132, 597, 599
- Axiom of choice, 392
- Axioms of probability measure, 16
- Azuma’s inequality, 647
- \mathfrak{B} , *see* Baire class
- \mathcal{B} , *see* Borel σ -algebra
- $\overline{\mathcal{B}}$, *see* Extended σ -algebra
- $b_n(k)$, $b_n(k; p)$, *see* Binomial distribution
- Bachelier, L., 349
- Backoff time, 277
- Baire
- class, 411
 - function, 409–411, 418, 470, 471, 490
- Baldi, P., 518n
- Ball in n dimensions, 295, 492–494, 498–500, 503, 504, 684
- Ballot
- problem, 44, 66
 - theorem, 46, 247
- Banach space, 484
- Banach’s match boxes, 261, 273
- Barbour, A. D., 668n, 678, 679n, 681n, 684n
- Bartholdi, J. J., 642n, 645, 646
- Base of number representation, 156, 158, 159, 378
- Basel problem, 123n, 594
- Bayes optimal decision, 66
- Bayes’s rule for events, 56, 66
- Bayesian prior, 62, 306
- Beardwood, J., 649n
- Beardwood–Halton–Hammersley theorem, 649
- Bell

- curve, 163, 164
 number, 653
 recurrence, 653
Bell, R., 606n
Benford, F., 290n
 Bennett's inequality, 647
 Bent
 binary digit, 150, 161, 162
 coin, 20, 184, 196, 235, 425
Berend, D., 653n
Berkeley, G., 772
Berlekamp, E., 131
 Bernoulli
 model of diffusion, 68
 schema, 235–274
 trials, 20, 235–237, 402, 420,
 425, 430, 433, 442, 508,
 510, 520, 524, 525, 551,
 552, 559, 568, 572, 583,
 595, 611, 614, 616, 620,
 646, 661, 662, 718, 720,
 746, 763
Bernoulli, D., 585
Bernoulli, J., 235
Bernoulli, N., 585, 586, 677
 Bernstein polynomial, 162,
 574, 604, 710
 in two dimensions, 605
 Bernstein's inequality, 647
Bernstein, S. N., 75, 76, 87,
 162, 532, 533, 572n, 605,
 709n, 711n
Berry, A. C., 724n
 Berry–Esseen theorem, 725,
 730, 736, 737
 Bessel function of the first
 kind, 515, 516
 modified, 363, 510
 Beta
 density, 305–307
 distribution, 307
 function, 305
 Bilateral exponential density,
 510
Billingsley, P., 695n
 Bin-packing problem, 394,
 633–637, 649
 Binary
 channel, 91
 digits, 141–144, 146, 148–
 150, 153, 177, 178, 496,
 502, 594
 entropy function, 612
 expansion, *see* Dyadic expansion
 integer programming, 751
 Binning number, 395, 634,
 649
 Binomial
 coefficient, 30, 51, 153, 237,
 574
 distribution, 20, 54, 95, 118,
 119, 150, 152, 162, 190–
 193, 195, 196, 237–239,
 256, 271, 393, 402, 430,
 432, 510, 519, 524, 525,
 562, 662, 689, 754, 763
 tail bound, 155, 611, 615
 theorem, 30
 Birth–death chain, 54–55
 Birthday paradox, 31, 452
 Bit (binary information digit), 502
 Bivariate normal density,
 215, 220, 228, 326, 338,
 453
Fischer, B., 349
 Black–Scholes formula, 349
 Blind
 equalisation, 346
 source separation, 346
 Bluetooth protocols, 138
Blum, R., 768n
Blumer, A., 604n
Bolthausen, E., 737n
 Bolzano's theorem, 344, 477
 Bolzano–Weierstrass theorem, 705, 706, 774, 781
 proof, 776
 Bonferroni's inequalities, 115, 115, 117
 Boole's
 inequality, 106, 111, 113,
 115, 122, 123, 125, 130,
 193, 195, 362, 382, 511,
 564, 565, 586, 589, 590,
 598, 603, 635, 647
 sieve, 125, 131
Boppana, R., 688n
Borchardt, C. W., 106
 Borel
 measurable function, 396,
 410, 411
 set, 23, 24–26, 27–30, 370,
 371, 374, 375, 384, 385,
 389, 390, 392, 395–399,
 406, 410–418, 420, 422,
 498–500, 504
 set (on the sphere), 768
 σ-algebra, 24–26, 385, 389,
 390, 396, 406, 412, 486
 Borel's law of normal numbers, 159, 394, 408, 714,
 744
Borel, E., 139, 140, 146, 150,
 155, 156, 159, 768n
 Borel–Cantelli lemma, xxiii,
 108, 262, 263, 362, 409,
 588, 599, 600
Bose, R. C., 135n
 Bose–Einstein distribution, 5,
 31, 136
 Boston Celtics, 33
 Boundary of a set, 343
 Bounded
 function, 779
 operator, 700, 701
 rectangle, 29
 sequence, 771
Bourbaki, N., xxi
Box, G. E. P., 426n
 Box–Muller construction, 426
 Branching process, 558
 Bridge hand, 9, 10
 balanced, 38, 39
 high pairs, 137
 trump distribution, 66
 voids, 270
 Broadcast authentication, 254
Brown, R., 348
 Brownian
 bridge, 366
 motion, 348–363, 366, 768
 Brun's sieve, 117
 Buffon's
 cross, 311
 needle problem, 291, 311
Buffon, G.-L., 292
 Burt's IQ data, 763–767
Burt, C., 763n, 765
 Busy period, 543

 C, *see* Cantor set
 C, C_b, Č_b (families of continuous functions), 692
 χ(G), *see* Chromatic number
 c.f., *see* Characteristic function
 California Institute of Technology, xvii

- Call arrivals, 297
 Campbell's theorem, 718
 Cantelli's strong law of large numbers, 605
Cantelli, F. P., *xvi*, 595n
 Cantor
 diagonalisation, 377, 377, 391, 391, 420, 706
 distribution, 404–406, 425, 434, 435, 444, 454, 463, 508
 set, 375, 377–380, 404, 425, 434
Cantor, G., 706
 Carathéodory's extension theorem, 371, 374, 375, 391, 395, 398, 413, 414
Carathéodory, C., 371, 372, 374, 385
 card, *see* Cardinality of a set
Cardano, G., 3
 Cardinal
 arithmetic, 375, 392
 number, 375
 Cardinality of a set, 20, 91
Carleson, L., 173
 Cartesian product, 212, 486
 Catalan number, 274
Catalan, E., 274
Catcheside, D. G., 137n
 Cauchy
 density, 208, 308, 311
 density in \mathbb{R}^3 , 313
 density in the plane, 313
 distribution, 309, 767
 sequence, 408, 409, 593, 594, 597, 599, 617, 708, 772, 776
 sequence in L^p , 484
 sequence in L^2 , 521, 785, 787, 788
Cauchy, A. L., 773
 Cauchy–Schwarz inequality, 483, 485n, 520, 521, 623, 635, 645, 675, 735, 736, 784, 785, 788
 proof, 787
 Cayley's formula, 102–106, 122
Cayley, A., 106
 CDMA, *see* Code division multiple access
 Central limit theorem, *xvi*, 163, 196, 350, 624, 704, 721, 733, 743, 747, 750, 752
 limit theorem for dependent summands, 751
 limit theorem for exchangeable variables, 768
 limit theorem for identical distributions, 720
 limit theorem for inversions, 767
 limit theorem for many dimensions, 748, 751, 753, 755–757
 limit theorem for order statistics, 314
 limit theorem for records, 767
 limit theorem for runs, 767
 limit theorem for sample median, 314
 limit theorem for triangular arrays, 738
 moment, 472
 tendency, 427
 term of the binomial, 196, 496, 556
 term of the multinomial, 271
 term of the trinomial, 272
 Centre of mass, 208, 430
 Centring, 207
 Certain event, 14
 Cesàro mean, 592, 725
 ch, *see* Convex hull
 Chain rule for conditional probabilities, 38, 126, 129, 688
Chandrasekhar, S., 287n
 Channel
 capacity, 506, 648
 equalisation, 346, 514
 matrix, *see* Mixing transformation
 noisy, 64, 65
 Characteristic
 exponent, 724, 767
 function, 171, 560, 698, 700, 718, 725, 726
 polynomial, 321–323
 set, 487
 Chebyshev's
 exponential bound, 630
 inequality, 153–155, 160, 162, 177, 242, 365, 462, 502, 503, 528, 563, 564, 565, 569, 570, 572, 573, 583, 584, 586, 588–590, 596, 599, 604–606, 615
Chebyshev, P., 153–155, 563n, 579n, 712
Chen, L. H. Y., *xvi*, 652n, 663, 677n
 Chernoff's inequality, 610, 612, 613, 615, 658, 659
 is exponentially tight, 612
Chernoff, H., 611n, 768n
Chervonenkis, A. Ya., *xvi*, 599n
Chevalier de Méré, 261
 Chi-squared
 density, 319, 333, 502, 752, 753, 759, 760
 distribution, 334, 759
 test, 759–763, 765, 766
 Chromatic number, 394
 Chromosome
 breakage and repair, 32, 256
 matching, 137
Chung, K. L., *xix*, 107, 252, 307
 Circular scan statistic, 686
 cl, *see* Closure of a set
 Clique, 138, 571, 604, 666–668
 Closed
 disc, 602, 607
 rectangle, 29, 607
 set, 26, 30, 378
 Closure of a set, 29, 712
 Coarser σ -algebra, 464, 485
 Code book, 501, 504–506, 519, 614–616, 647
 Code division multiple access, 138, 192, 193n
 Cognitive radio, 69
 College admissions, 40, 69
 Collisions in transmissions, 138, 277
 Column sum, 674
 Combinatorial
 lemma, 548
 probability, 20
 Compact set, 708, 714
 Compactification of real line, 692
 Complete
 graph, 109, 138
 inner product space, 485, 521, 785

- measure, 374, 392
normed space, 484
Completely monotone function, 532, 533, 711
Completeness axiom, 771, 775
Completion of measurable space, 373–375, 392, 486
Complex random variable, 445
Component
 of a graph, 121, 681
 of \mathbb{R}^n , 343, 607
Compound Poisson distribution, 538, 558
Computational neuroscience, 193
Concentration function, 618, 624, 633, 648, 649
 of measure, xvii, 609, 617, 625, 647
Concept (in machine learning), 603
Conditional density, 219, 219
distribution, 218, 220
expectation, 220, 221, 463, 464, 485, 521, 522
probability, 35–70
variance, 220
Conditionally independent events, 72, 74, 75, 90
Confidence
 in estimate, 63, 241, 242, 567, 603, 616
 interval, 63, 331, 754, 759, 766
Configuration function, 650
Conjugate exponent, 482
Connected graph, 102, 121–125, 137, 681
Continuity
 from below, 373
 from above, 373
 of measure, 16, 19, 21–23, 38, 89, 106, 108, 293, 390, 391, 398, 408, 409, 441, 714
Continuous
 distribution, 293, 402–406, 518
 everywhere, differentiable nowhere, *see* Nowhere
 differentiable function from the left, *see* Left continuous function
 from the right, *see* Right continuous function
 function, 777, 778
 sample space, 11, 23–24
Convention
 for d-adic expansion, 158
 for binary representations, 22, 141
 for binomial coefficients, 30, 237, 246
 for conditional expectation, 220
 for convolutional operators, 696
 for countable operations, 15
 for cross-references, xxiii
 for d.f.s and induced measures, 399
 for densities, 207
 for derivatives, 532, 710
 for difference operators, 574
 for distributions, 198
 for Fourier transforms, 166
 for intersections and unions, 94
 for Kullback–Leibler divergence, 480
 for ladder index distributions, 553
 for logarithms, 30, 575
 for normalisation, 17
 for positive parts, 279
 for random walks, 543
 for renewal process, 534, 535
 for repeated convolutions, 535
 for runs, 559
 for sets, 12, 13, 127
 for terminology, xxiii
 for vectors, 216, 320, 359
Convergence
 a.e., 158, 162, 407–409, 460, 462, 646, 649, 785, 787, 788
 bounded, 487, 487
 improper, 705
 in distribution, 691, 705, 717, 721, 742, 745, 748, 751, 752, 767, 768
 in law, 691
 in L^2 , *see* Convergence in mean-square
 in mean-square, 354, 594, 785
 in measure, 409
 in probability, 155, 162, 409, 598, 646
 of sequence, 771
 of series, 147, 289, 423, 593–595, 709, 711
 pointwise, 774
 proper, 705, 714
 uniform, 147, 401, 574, 596, 598, 600, 697, 707, 710, 717, 722, 739, 750, 775, 781
 vague, 691, 695, 696, 700, 702, 705, 710, 711, 714, 717, 718, 721, 724, 725
 weak, 695
 with probability one, *see* Convergence a.e.
Convex
 combination, 478, 621–623, 626, 702
 distance, *see* Talagrand's convex distance
 function, 478, 520, 630, 632, 650
 hull, 621, 621, 622, 623, 627
 polyhedron, 504
 set, 607
Convolution
 is smoothing, 507, 512, 536
 of arithmetic variables, 508
 of Cauchy densities, 309, 311
 of continuous variables, 508
 of distributions (densities), 203, 204, 230, 507, 506–511, 696, 703, 715, 721, 740
 of exponential densities, 298, 314
 of functions, 167, 512, 767, 781
 of gamma densities, 304
 of ladder epoch distribution, 551
 of normal distributions, 318

- of Poisson distributions, 256, 528
of uniform densities, 279, 284
property of characteristic functions, 560, 726
property of transforms, 152, 167, 170, 513, 515, 525, 530
Convolutional kernel, 512, 513, 725
operator, 695–698, 702–704, 714, 725
smoothing, 698, 725
Coordinate transformation, 409–411, 413
Copernican Principle, 61–64
Correlation coefficient, 474
definition, 213, 473
function, 341
inequality, 520, 687
Countable cover, 384
set, 14, 19, 27, 375, 376, 391
Coupling, 669, 670, 671, 673, 675, 679, 680, 682, 684, 686
Coupon collector’s problem, 115, 119, 259, 742–745
Covariance definition, 213, 473
matrix, 217, 360, 748, 751, 752, 756–758
Cover of a set, 774, 776
of circle by arcs, 283, 284, 681
Cover, T. M., xxiv, 273, 606n
Covering problem, 283
Cramér’s estimate for ruin, 542
Cramér, H., 542n
Cramér–Rao bound, 521
Craps, 6, 81, 201, 557
Crofton’s method of perturbations, 296
Crofton, M. W., 295
Crystallography, 520
Curve (inside unit square), 638
Cycle in graph, 102, 394
in permutation, 745–747
Cyclical arrangement, 547–549, 551, 552
rearrangement, 745, 746
Cylinder set, 60, 148, 149
 Δ , *see* Difference operator, successive differences, triangular function
d.f., *see* Distribution function
Darmois, G., 344n
Davis, P. J., xix
de Finetti’s theorem, 282–284, 679
de Finetti, B., 283n
de la Vallée Poussin, C.-J., 102, 579
de Moivre’s theorem, 163, 183, 184, 189, 190, 195, 196, 633, 721
de Moivre, A., xvi, 163, 184, 719
de Moivre–Laplace theorem, 205, 314, 577, 690, 720, 755
de Montmort, P. R., xvi, 49, 98, 585, 651, 672, 677n, 685, 686
de Morgan’s laws, 13, 33, 88, 89
Decision rule, 65, 66, 504–506, 519
minimum distance, 519
Decorrelating transformation, 327, 329
Decreasing family of sequences, 687
sequence of numbers, 771
sequence of sets, 34
Defective distribution, 407, 540, 541, 705
random variable, 407, 705
Degenerate distribution, 394, 401, 480, 481, 756
normal, 330
random variable, 430
Degree of a vertex, 102, 126, 394
Degrees of freedom, 331, 333, 335
Delta function, 173
Denial of service attack, 233, 254, 761
Dense set, 705, 714, 716
Density, 24, 206, 403
of a number, 582
of digits, 156, 160, 394
of stars, 296
on Euclidean plane, 211
Dependence penalty, 664
Dependency graph, 126, 664, 668
Dictionary of translation, 149, 150, 153, 177
Dido of Tyre, 496
Difference of sets, 12
operator, 343, 573, 709
Differentiation operator, 167
property of transforms, 167, 171, 172, 526, 530
Difficulties with the boundary, 683, 685n
Direct sequence spread spectrum, 138n, 192
Dirichlet density, 313
Discrepancy, 426, 559, 725, 726, 729, 731, 733, 735, 736
Discrete distribution, 20, 197, 399–402
sample space, 9, 19–23, 197
Disjoint cover, 384
events, 14
sets, 12
Distribution function, 206, 370, 371–372, 374, 375, 392, 398–406
in rôle of function or measure, 399
obtained by randomisation, 221, 232
of a measure, 375, 530
of arithmetic variable, 198
of spacings, 281
Dobinski’s formula, 653, 654
Dominated convergence theorem, 460, 461, 462, 464, 487, 592, 595, 733, 774
random variable, 433, 437, 438, 461, 462
Domination condition, 461
DoS, *see* Denial of service

- Double sample, 426, 596, 608
Dudley, R. M., 562
 Dyadic
 approximation, 263
 expansion, 21, 141, 142,
 150, 156, 394, 420, 717
 interval, 352
Dynkin, E. B., 87, 88, 92
- Eagleson, G. K.*, 678n
 Edge of a graph, 85, 102, 394,
 650
 Ehrenfest model of diffusion,
 52–55, 68
Ehrenfest, P. and T., 52n
Ehrenfeucht, A., 604n
 Eigenfunction of Fourier operator, 172
 Eigenvalue, 322, 323, 332
 equation, 323, 324, 339
 Eigenvector, 322, 323, 332
Einstein, A., 348, 350
 Ellipsoid in n dimensions,
 329
 Empirical
 distribution, 241, 426, 559,
 595–598
 estimate of π , 291
 frequency, 599, 600
 process theory, xvi
 Empty set, 12
 Energy, 518, 784, 786
 Equicontinuous family of functions, 697, 707, 716, 780
 Equivalence
 class, 392, 484
 lemma, 441, 460
 theorem, 693, 696, 697, 702,
 708, 711, 714, 718, 768
 Equivalent
 random variables, 484, 520
 sequences, 591
 sets, 12
Erdős, P., 86n, 110, 111n, 119,
 121n, 123, 128n, 134n,
 307
 erf, *see* Error function
Eriksson, J., 476n
Erlang, A. K., 304
 Erlangian density, 304
 Error
 function, 316
 in estimate, 241, 242, 567,
 603
 prediction, 518
 probability, 615
Esseen, C-G., 724n, 730n
 Estimator
 empirical, 226
 least-squares, 222, 223, 518
 linear, 222, 223, 518
 maximum likelihood, 614
 Euler square, 134
 Euler's officer problem, 134
Euler, L., 123n, 144, 157, 523,
 594
 Evolution of mosquito populations, 285
 Exchangeable random variables, 281, 283, 668, 671,
 672, 684, 733, 768
 Excluded interval, 713
 Expansion of a number in a given base, 158
 Expectation, 465–471, 649
 and median, 646
 change of variable, 472,
 473
 definition, 199, 207, 427,
 430, 437, 445
 identity, 232, 463
 is additive, 466
 is homogeneous, 429, 431,
 448, 462
 is linear, 432, 448, 464, 472,
 480
 is monotone, 429, 433, 451,
 460, 462, 464, 480, 561,
 565, 567, 694
 notation, 429, 469, 471
 Exponential
 density, 24, 205, 275, 293,
 297, 298, 301, 303, 319,
 403, 510, 721
 distribution, 292–294, 297,
 298, 312–314, 403, 450,
 524, 543, 546, 686, 717
 generating function, 653
 Extended
 real line, 372, 406
 σ -algebra, 406
 Extension of measure, 374,
 385
 Extreme value distribution, 233
 \mathcal{F} -measurable, *see* Measurable with respect to a σ -algebra
 $\mathfrak{F}, \mathfrak{G}, \mathfrak{H}$: notation for convolutional operators, 696
F, *see* Measure, induced by a distribution
F*, *see* Outer measure
F-density, 337
F-statistic, 336, 337
 Factorisation property, 464
 Fading in wireless communication, 64, 319
 Failure
 in Bernoulli trial, 236
 run, 559
 Fair game, 258, 273, 434, 585,
 606
 Falling factorial, 51, 84, 655,
 659, 666
 Family of half-closed intervals, 15, 371
Fan, C. T., 233n
Fang, S. C., xxiii, 750n
Fatemeh, O., 233n
 Fatou's lemma, 459, 460, 464,
 484, 787
 Fattening (of a set), 618, 621,
 624
 Fejér kernel, 725, 727
Feller, W., 5n, 271, 288–290,
 307, 542n, 547n, 586n,
 606n, 696, 720n, 730n,
 738
Fermat, P., 261, 273
Fermi, E., 567
 Fermi–Dirac distribution, 5,
 32
 Fiancée problem, *see* Marriage problem
 Field, *see* Algebra of events
Fielding, H., xv
 Filtration, 522
 Finer σ -algebra, 464
 Finite
 alphabet, 64
 cover (subcover), 774, 776
 measure, 372, 470, 487, 488
 projective plane, 134
 Finite-dimensional distribution, 340, 341
 Fisher information, 520
Fisher, R. A., 241, 269, 331n,
 336n
 FKG inequality, 520
Flood, M. M., 69
 Fluctuation theory, 33, 250

- 305, 307, 550–555
 Flush (poker), 31
 Fluxion, 772
 Fock space, 768
Fortuin, C. M., 520n
 Four-of-a-kind (poker), 31
 Fourier
 basis, 786
 inversion, 172, 175, 176, 178, 699, 726, 727
 series, 196
 transform, 166, 176, 178, 181, 309, 512–514, 560, 698, 699, 725–727
 transform in two dimensions, 514
 transform of unit disc, 515
Fourier, J. B. J., 166, 572
 Fourier–Lebesgue transform, 560
 Fourth moment, 588, 605, 743, 744
 Fractal, 346, 347
 Fractional part
 of random variable, 288
 of real number, 196, 715
Franklin, J., xxiii
 Free energy, 650
 Frequency-hopping spread spectrum, 138
 Fubini’s theorem, 489, 489, 490, 492, 507, 517, 531, 596, 598, 628, 629, 699, 703, 735
 fails for outer integrals, 518
 Full house (poker), 31
 Function
 is concentrated on, 207
 space, 608
 Fundamental theorem
 of algebra, 321, 323
 of arithmetic, 582
 of expectation, 471, 472

 $g_V(x; \alpha)$, see Gamma density
 $G_{n,p}$, see Random graph
Gabor, D., 288
Gale, D., 69n
 Galton’s height data, 223–227, 766, 768
Galton, F., 220, 223, 225n
 Gambler’s ruin, 68
 Gamma
 density, 298, 298, 300, 301, 303–305, 313, 319, 334, 721
 distribution, 298, 302, 525, 721
 function, 195, 303, 493–495
 Gaps
 between spacings, 679–681
 in the Poisson process, 536–538
Gauss, C. F., 315, 456n
 Gaussian
 channel, 501, 505, 506, 616, 647
 characteristic function, 171, 726
 density, see Normal density
 impulsive family, 175
 inversion formula, 172, 175
 isoperimetry, 498–500, 505
 kernel, 698
 noise, 501, 502, 647
 process, 341, 357, 366
Gedanken experiment, 4, 9, 11, 21
 General position, 607
 Generating function
 definition, 524, 610, 710
 of arithmetic distributions, 557
 of ladder index, 553, 554
 of renewals, 535
 of return to origin, 555
 Geodesic metric, 683n, 768
 Geometric
 distribution, 20, 257, 258–260, 262, 272, 275, 313, 427, 451, 510, 524, 525, 558, 721, 742, 743
 mean, 481
 Gilbert channel, 92
Gilbert, E. N., 92n, 682n
Gilovich, T., 33n
Ginibre, J., 520n
 Girth of a graph, 394
Glivenko, V. I., xvi, 595n
 Glivenko–Cantelli theorem, 596, 599, 600, 602, 603, 608
Gnedenko, B. V., 559n
Goldstein, R. M., 338n, 341
 Goncharov’s theorem, 747
Goncharov, V., 747n
 Gosset, W. S. (Student), 335n
Gott, J. R., 63n
Gould, S. H., 3n
 Grading on a curve, 316
 Graeco-Latin square, 134
 Graph, 102, 132, 394
 colouring, 394
Gray, J. B., 762n
 Greatest lower bound, *see Infimum*
Greenspan, A., 350
Gromov, M., xvii, 617
 Gromov–Milman formulation, 616–619, 624, 633
 Growth function, 600, 601, 607, 608
Gunter, C., 233n, 254n, 761n
Gupta, P., 684n

 H_0 , *see* Heaviside function
 \mathfrak{H}_0 , *see* Heaviside operator
 Haar system of wavelets, 352, 353–355, 359, 786
Hadamard, J., 102, 579
 Half-closed rectangle, 29
 Half-space, 602, 607
 affine, 608
Hall, P., 682n
Halmos, P., 89, 92
Halton, J. H., 649n
Hammersley, J. M., 649n
 Hamming
 distance, 614, 619, 622–624, 633, 634, 637, 644, 645, 648, 649
 error vector, 620, 621, 627n, 634
 Hard-limited random waveform, 338, 341
 Hardy’s law, 81
Hardy, G. H., 80n, 346, 485n, 581, 582n, 712, 715n
 Hardy–Ramanujan normal order theorem, 583
 Harker–Kasper inequality, 520
 Harmonic series, 519, 556, 594
 Hat check problem, 47, 452
Haussler, D., 604n
Hearnshaw, L., 767n
 Heat equation, 572
 Heaviside
 distribution, 401, 534, 697, 700, 703, 721, 725

- function, 400, 401, 410, 528, 561, 609, 778
operator (\mathfrak{H}_0), 697
- Heaviside, O.*, 779
- Heavy-tailed density, 308
- Hebb, D. O.*, 194n
- Hebbian learning, 194, 195
- Heine–Borel theorem, 383, 774
proof, 776
- Helly's selection principle, 533, 705, 708–711, 713, 714, 717, 718
- Helly, E.*, 705n
- Helsinki University of Technology, xvii
- Herault, J.*, 345n
- Hermite, C.*, 347
- Hilbert
curve, 638–643
space, 485
- Hilbert, D.*, 347, 638n
- Hildebrand, D. K.*, 762n
- Hoare, C. A. R.*, 459n
- Hoeffding's inequality, 598, 612, 616, 620, 621, 625, 635, 647
for the binomial, 562, 647
- Hoeffding, W.*, 612n
- Hölder's inequality, 483, 484, 628, 631, 725, 735, 736, 743, 787
- Holst, L.*, 668n, 679n, 681n, 684n
- Hopfield, J. J.*, 193n
- Horne, E.*, xxiv
- Horse race, 606
- Hot hand in basketball, 32, 33
- Hotelling's theorem, 365
- Huygens's principle, 309
- Huygens, H. C.*, 310
- Hypergeometric
distribution, 52, 268–271, 273, 599, 604
tail bound, 604, 608
- Hyperplane, 504, 607, 608
- Hypothesis (in machine learning), 603
- \mathcal{I} , *see* Family of half-closed intervals
- i.o., *see* Infinitely often
- $I_0(\cdot)$, $I_k(\cdot)$, *see* Bessel function of the first kind, modified
- Identity map, 375, 396, 745
- Imbedded product measure, 420
- Importance
density, 604
sampling, 604
- Impossible event, 14
- Impulsive family of functions, 173, 175
- Incident edge, 102
- Inclusion-exclusion, 93–101, 105, 113, 115–117, 120, 124, 136, 283, 413, 452, 673, 675
- Increasing
family of sequences, 687
sequence of numbers, 771
sequence of sets, 34
- Increment, 314, 360, 362
- Independence, 71–92, 415–422
number of graph, 394, 650
set of vertices (events), 126, 664, 666, 668, 688
sieve, 125, 687
- Independent
binary digits, 144–148, 150, 151, 161, 178, 420, 594
events, 37, 71–73, 83, 203
families of events, 87–90
increments, 302, 350, 360–362
random variables, 192, 202, 202, 212, 217, 415, 418, 419, 464, 490, 560, 734
sampling, 418, 602, 649, 679
selection, 684
 σ -algebras, 89, 419
trials, 5, 17, 81–87, 192, 194, 196, 202, 661
- Indicator
family, 660, 669, 674
function, 33, 151, 153, 155, 489, 492
random variable, 240, 402, 430, 518, 589, 657
- Induction method, *see* Lagrange's induction method
- Inequality of arithmetic and geometric means, 481, 482, 483, 520, 749
- Infimum, 27, 385, 386, 405, 773
- Infinite product
identity, 141, 144, 147
space, 87, 422
- Infinite-dimensional sphere, 768
- Infinitely often, 107, 252, 263, 423, 424
- Information rate, 616
- Inner
envelope, 424, 426
product in L^2 , 485
product of Euclidean vectors, 192, 320, 340, 675
product of functions, 353, 784, 787
- Inspection paradox, 299, 546
- Integer part of real number, 196
- Integrable
function, 166
random variable, 437, 447, 482
- Integration, 465–471
by parts, 491, 540
- Inter-arrival times, 298, 534, 542
- Interchange in order of derivative and integral, 168, 230, 286, 526, 698, 781
derivative and sum, 272, 710
expectation and sum, 355, 448
integrals, 151, 173, 176, 178, 221, 485–492, 699, 783
limit and integral, 147, 148, 180, 181
limit and sum, 441
sum and integral, 157, 300
sums, 553, 591
- Interior
of a set, 343
point, 26
- Intersection of sets, 12
- Interval of continuity, 691
- Intervals are uncountable, 377
- Invariant
distribution, *see* Stationary distribution

- subspace, 321–323, 332
 Inversion theorem
 for characteristic functions, 700
 for Fourier transforms, 175
 for Laplace transforms, 527, 530, 653
 Inversions in permutation, 767
 Ising spin glass, 518
 Isolated
 point in geometric graph, 682, 684, 686
 vertex in a graph, 120
 Isoperimetric problem, 497, 500
 theorem, 482, 498–500, 504, 505, 624
 Iterated integrals, 485–492, 506, 517

 $J_0(\cdot)$, $J_1(\cdot)$, $J_k(\cdot)$, *see* Bessel function of the first kind
 Jacobian, 228, 229, 313, 469
 Janson's inequality, 688
Janson, S., 668n, 679n, 684n, 688n
Jaynes, E. T., 306
 Jensen's inequality, 480, 481, 519, 520, 522, 613, 628
Jones, A., xxiv
 Jump
 discontinuity, 400, 424
 size, 400, 401
Jutton, C., 345n

 K_4 , K_j , K_k , K_n , *see* Complete graph, clique
Kac, M., 87, 139n, 307
Kahane, J.-P., 363n
Kahneman, D., 33n
Kamiń, L., 765
 Kantorovich's inequality, 520
Kasbekar, G., xxiii
Kasteleyn, P. W., 520n
Khan, F., 233n
Khanna, S., 233n, 254n
Khinchin, A. Ya., 563n
 Kleitman's lemma, 687
Kleitman, D. J., 687n
 Knox's leukaemia data, 678
Knox, G., 678n
Knuth, D., xxi, 459n, 568n
 Kolmogorov's criterion, 590, 592

 maximal inequality, 588, 590, 593
Kolmogorov, A. N., 4n, 9, 16, 87, 363, 422, 423, 588n, 591, 595, 605
 Kolmogorov-Smirnov test, 759
Körner, T. W., 173, 184, 369, 714, 753–755, 765–767
Korolev, V. Yu., 730n
Koroljuk, V. S., 559n
 Kronecker's theorem, 712, 713, 714, 716
Kronecker, L., 712n
 Kullback–Leibler divergence, 480, 481, 521, 611, 646
Kumar, P. R., 684n
Kunniyur, S., 685n

 \mathcal{L} : notation for probability law, *see* Distribution
 L^2 -space, *see* Space of square-integrable functions
 L^p -space, 482–485
 λ : notation for measure, *see* Lebesgue measure
 λ -class, 88, 89, 390, 487, 488
 Labelled sample, 602
Labeyrie, A., 511n, 513, 516
 Ladder
 epoch, 544
 index, 426, 546–550, 551–554, 559, 588
Lagrange, J. L., 523
Lamarr, H., 138n
 Laplace transform, 523–532, 712, 718
 definition, 523, 530, 710
 determines moments, 526
 is completely monotone, 711
 of common distributions, 524
 Laplace's
 formula, 196
 law of succession, 59–62, 69, 72, 75, 275, 307
 method of integration, 187, 494
 theorem, 196
Laplace, P. S., 5, 24, 61, 184, 196, 275, 523
 Laplace–Stieltjes transform, *see* Laplace transform

 Laplacian uncertainty, 276
 Large deviation theorem, 191, 192, 195, 196, 562, 612, 750
 Latin
 square, 134
 transversal, 131–135
 Latitude, 214, 229, 285
 Lattice
 distribution, 200–204
 point, 200
 random variable, 198, 200
 Law
 of cosines, 340
 of small numbers, 253
 Le problème des ménages, 137, 677, 686
 rencontres, 47, 98, 119, 137, 668, 672, 686
Lea, D. E., 137n
 Leaf of a tree, 103
 Least upper bound, *see* Supremum
 Lebesgue
 integral, 465–471, 518
 measurable set, 375, 392
 measure, 144, 145, 149, 150, 155, 156, 264, 276, 375–380, 392, 403, 420, 422, 471, 491, 517, 717
 Lebesgue's decomposition theorem, 425
Lebesgue, H., 23, 347, 429, 436, 467
 Lebesgue–Stieltjes integral, *see* Lebesgue integral
Ledoux, M., 624n
 Left continuous function, definition, 778
Leibniz, G. W., 24, 465, 772
 Lemma of Cesàro means, 592
 Length
 of interval (Borel set), *see* Lebesgue measure
 of random chain, 517
 Let's make a deal, 42
 Level set, 404, 405, 433–435
 Levi's theorem, 157, 462, 486
Levi, B., 157, 439
 Lévy sandwich, 181, 182, 358, 693, 696, 709, 714
Lévy, P., 181, 250n, 307, 351, 498, 693, 720, 724n, 750
Lewis, S., xxiv

- Liapounov condition, 741,
 744, 767
Liapounov, A. M., 741
 Lightning strikes twice, 277
 lim inf, *see* Limit inferior
 lim sup, *see* Limit superior
 Limit inferior
 of a real sequence, 774
 of a sequence of sets, 107,
 408
 Limit superior
 of a real sequence, 774
 of a sequence of sets, 107,
 408
 Lindeberg condition, 738,
 739–741, 747, 767
Lindeberg, J. W., 184, 720n,
 723, 750
 Linear congruential generator, 568
 Linear transform, 227, 321
 diagonalisation, 324, 329,
 339, 365
 for sample variance, 332
 of normal variables, 326,
 359, 360
 positive definite, 324, 329,
 339
 symmetric, 322, 323
 Linearly ordered set, 455
 Lipschitz
 condition, 617, 633, 634,
 636, 643, 644, 649
 constant, 618, 619
 function, 162, 618, 618, 640,
 642, 649
Littlewood, J. E., 485n, 712
 Local
 limit theorem, 184, 189–
 191, 195, 196
 method (of Poisson approxima-
 tion), 663–668,
 684n
 Log convex function, 272
 Log sum inequality, 519, 646
 Logarithms, *see* Convention
 for
 Lognormal distribution, 558
 Longest
 common subsequence, 650
 increasing subsequence,
 636–638
 Longitude, 214, 229, 285
 Lord Rayleigh's random
- flights, 284–288, 752
 Lottery, 31
 Lovász local lemma, xvi, 125,
 128, 129–132
Lovász, L., 127, 128n
 Lower derivative, 361
 L^p -space, 482
Lucas, E., 137, 677, 686
Luczak, T., 688n
- $\bar{\mu}$, *see* Extension of measure
 Machine learning, xvi, 603
Mandelbrot, B., 346n
 MAP, *see* Maximum a poste-
 riori probability
 Marginal
 density (distribution), 201,
 202, 203, 212, 220, 391,
 415–418, 506, 507, 521
 of normal density, 325
 Markov property, 294
 Markov's
 inequality, 561, 562
 method, 177–181, 196, 287
Markov, A., 181, 750
 Marriage problem, 69, 273
 Martingale, 522
 bounded-difference, 647
 transform, 522
 Matched filter receiver, 192
 Matchings (le problème des
 rencontres), 49, 685
 multiple, 685
 Maximal inequality, 588
 Maximum
 a posteriori probability, 59,
 66
 likelihood principle, 241,
 269, 647
 of normal variables, 646
 of random walk, 248, 274,
 554
 Maxwell's distribution of ve-
 locities, 320, 752, 753
 Maxwell–Boltzmann distri-
 bution, 31
McCarthy, J., 366n
McCulloch, W. S., 193n
 McCulloch–Pitts neuron, 193
McEliece, R. J., 195n
 Meals-on-Wheels, 645
 Mean
 definition, 199, 207, 212,
 427, 472
- of binomial, 239
 of exponential, 293
 of geometric, 258
 of negative binomial, 260
 of normal, 316
 of Poisson, 255
 of uniform, 276
 recurrence time of runs,
 560
 vector, 217
 Mean-square approximation
 theorem, 354, 786
 proof, 788
 Measurability theorem, 396,
 397, 406, 407, 412
 Measurable
 function, 393–397
 partition, 42, 56, 430, 475
 set, 372–374
 space, 372, 393, 395
 with respect to a σ -algebra,
 395
 Measure
 definition, 372–375
 induced by a distribution,
 370, 390, 397–399
 on a ring, 380–384
 σ -finite, *see* σ -finite mea-
 sure
 Median, 428, 509, 510, 518,
 618, 636, 637, 645, 649,
 650
 Memoryless property, 259,
 275, 293
 of exponential distribu-
 tion, 293, 294, 303
 of geometric distribution,
 259, 260
 Mengoli's inequality, 519
Mengoli, P., 123n, 519, 594
 Merton, R. C., 349
 Method of
 coupling, 668–684
 images, 246, 274
 truncation, 564, 586
 Metric, 520, 617, 718, 781
Meyer, Y., 359n
Meyler, P., xxiv
Milman, V. D., xvii, 609, 617
 Minimal σ -algebra, 25
 Minimax decision, 70
 Minimum degree of a graph,
 394
 Minkowski's inequality, 483

- 485
 Mixed moment, 473
 Mixing transformation, 342, 344, 345
 Mixture distribution, 702
 Möbius function, 100, 101
 Mode, 428
 Modulus inequality, 447, 451, 483, 664, 702, 727, 728, 732
 Moment
 generating function, *see* Generating function, definition
 of inertia, 208
 Moments
 definition, 472, 517
 determine distribution, 529, 718
 of the normal, 165
 Monoid, 703
 Monotone
 class of sets, 88, 92
 class theorem, 92
 convergence theorem, 439, 442, 445, 448, 451, 459, 460, 464, 466, 475, 487–490, 522, 532, 655, 787
 property of graphs, 124
 Monotonically related indicator variables, 671, 671
 Monotonicity
 of measure, 18, 38, 106, 127, 129, 372, 376, 398, 416
 of outer measure, 386
 Monte Carlo method, 566–568
Moon, J. W., 106n
Moore, E. H., 347, 638
Moran, P. A. P., 271
 Mosaic process
 in a cube, 682
 in a square, 683
 on a circle, 681
 Mother wavelet, 353
 Moving truncation, 588, 591
Muller, M. E., 233n, 426n
 Multilevel sequences, 647
 Multinomial
 distribution, 271, 754, 755
 theorem, 154, 157
 Multiplicative identity, 703
 Multivariate normal density, 324–330, 359, 758
 Multivariate uniform distribution
 in the unit ball, 214
 in the unit disc, 214, 220
 in the unit square, 214
 on the sphere, 284
 Mutual information, 521
 Mutually exclusive events, 14, 90
 $\mathcal{N}(\mu, \sigma^2)$, *see* Normal distribution
 n-dimensional
 distribution, 414
 simplex, 283
 n-server queue, 314
 Naïve algorithm for sorting, 456
 Nearest-neighbour distance, 682
 Negative binomial
 coefficient, 51
 distribution, 20, 260, 273, 304, 313, 451, 525, 721
 Negative part of random variable, 437
 Negatively related indicator variables, 671, 671, 679, 680, 687
 Negligible set, 361
 Neighbouring vertices, *see* Adjacent vertices in a graph
 Neural computation, 518
 Neuron, 193
Newton, I., 140, 239, 772
 Non-measurable
 function, 464
 set, 392
 Norm
 L^2 , 485
 L^p , 482, 484, 520, 682
 of Euclidean vector, 321, 619, 620, 621, 623, 625, 630, 640, 682, 749, 755, 757
 of function, 353, 697, 701, 718, 723, 784
 of operator, 701, 701, 704, 717
 square is convex, 627, 630
 Normal
 approximation, 731, 733, 750
 density, 24, 164, 205, 315, 315–366, 404, 510
 density in n dimensions, 324, 328, 502
 distribution, 315, 315–366, 404, 516, 646, 690, 698, 699, 720, 723, 724, 731, 733, 740, 742, 744, 747, 748, 751, 752, 757, 758, 767, 768
 law, *see* Central limit theorem
 number, 155, 156, 159, 160
 order of integers, 582
 tail bound, 166, 193, 317, 355, 610, 646, 647, 732
 Normalisation
 constant, 165, 303, 305, 324
 of measure, 16, 19, 37, 54, 55, 390
 Normalised measure, 531
 Normed space, 484, 700–703
 Nowhere differentiable function, 346–348, 361, 366, 641
Nukpezah, J., xxiii
 Null
 hypothesis, 759
 set, 407, 593
 space, 332
 Number representation, *see* Expansion

 Ω , *see* Sample space
 ω , *see* Sample point
 Occupancy configuration, 31
 Ockham's razor, 754n, 760
Ockham, W., 24, 754n
 One pair (poker), 31
 Open
 ball, 29n
 cover, 383, 774, 776
 rectangle, 28, 691
 set, 26, 28, 30, 378, 602, 607
 Operator
 difference, 701
 product, 703
 scalar multiplication, 701
 sum, 701
 Optical holography, 288
 Optimal
 decision, 69
 stopping, 69
 tour, 643, 644, 646, 649

- Optimum receiver, 647
 Oracle, 602
 Order
 notation (\mathcal{O} , \asymp , \sim , \mathfrak{o}), 575
 of a component, 121
 statistics, 313, 313, 517, 642
Ore, O., 3n
 Orthant, 607
 Orthogonal
 functions, 143, 785
 Latin squares, 134
 projection, 521
 random variables, 522
 subspace, 756
 transformation, 324, 756, 768
 vectors, 321
 Orthonormal
 basis of eigenvectors, 323, 329, 332, 339, 756
 system of functions, 160, 353, 786
 Oscillation of a function, 694, 696, 780
Ott, R. L., 762n
 Outer
 envelope, 424, 426
 integral, 464, 518
 measure, 374, 385, 384–392, 416, 464
- Φ , $\Phi_{\mu, \sigma}$, *see* Normal distribution
 ϕ , $\phi_{\mu, \sigma}$, *see* Normal density
 π -class, 87, 89, 390, 417, 419, 486–488
 π - λ theorem, 87, 88, 89, 92, 390, 417, 419, 486, 487, 488
 $p(k; \lambda)$, *see* Poisson distribution
 PAC learning, *see* Probably approximately correct learning
 Packet-switched communications, 303
 Pairwise independence, 73, 76, 77, 91
Paley, R. E. A. C., 348n, 351
 Parallelogram law, 521
 Pareto
 density, 308
 principle, 308
Pareto, V., 308
- Parker, E. T.*, 135n
 Parseval's equation, 161, 176, 180, 352, 354, 786
 proof, 788
 Partial order, 520
 Partition, 137, 161
 function, 650
 Pascal's triangle, 30, 54, 114, 237, 260, 279, 450, 574, 602
Pascal, B., 261, 273
 Path in a graph, 102, 685
Peano, G., 347, 638n
Pearson, E. S., 335n
Pearson, K., 285n, 286, 319, 759n
Penrose, M., 685n
 Pepys's problem, 232, 238, 272
Pepys, S., 238
 Percolation, 650
 Permanent of a matrix, 136
 Permutation, 83–85, 91, 131–136, 642, 672, 674, 745
 matrix, 674–676
Perrin, J. B., 348, 350, 361
 Phase transition, 124, 193, 195, 570, 666, 668
 Pigeon-hole principle, 584, 649
Pinkham, R. S., 290
 Pinsker's inequality, 646
Pitts, W. H., 193n
 Plancherel's equation, 177
Platzman, L. K., 642n, 645
 Plucked string function, 366
Poincaré, H., 288, 291, 347, 349
 Point set topology, 26–30
 Poisson
 approximation, xvi, 113–119, 651–688, 690, 731, 733
 approximation to the binomial, 118, 253, 662
 characterisation, 654
 distribution, 20, 118, 119, 196, 255, 271, 272, 297, 302, 402, 442, 510, 524, 527, 528, 570, 652, 690, 767, 768
 ensembles of stars, 294
 paradigm, 113, 119, 121, 138, 570, 651
 process, 301, 302, 313, 536,
- 538, 542, 718
 sieve, 687
 tail probability, 658, 659
Poisson, S. D., xvi, 118, 253n
 Poissonisation, 649
 Poker, 9, 10, 31, 253, 272
 Polar coordinates, 229, 340, 363, 364
 Polish space, 599, 617
Pollaczek, F., 248, 545, 554
 Pollaczek–Khinchin
 formulæ, 301, 541, 542, 545, 559
 mean value formula, 546
 theorem, 541
Pollard, D., 599n
Pollock, J., 348
 Polls, 158, 239–244
 Pólya's
 theorem, 557
 urn scheme, 49–52, 68, 522
Pólya, G., 485n, 557n
 Polynomial approximation, 571
 Portfolio, 195, 606
 Positive
 measure, 381
 part of random variable, 437
 part of real number, 278
 random variable, 198, 207, 234, 437, 449, 523
 Positively related indicator variables, 671, 672, 673, 686
 Positivity of measure, 16, 18, 19, 22, 23, 37, 372
Posner, E. C., 195n
 Preimage of a set, 395
 Prime number theorem, 102, 579
 Primes are independent, 582
 Probabilistic method, 110
 Probability
 as expectation of indicator, 467
 density, *see* Density
 distribution, *see* Distribution
 law, *see* Distribution
 mass function, 198
 measure, 16–18
 of ruin, 540–542, 545

- sieves, 93–138, 652, 668, 685
space, 17
- Probably approximately correct learning, 604
- Problem of the points, 261, 273
- Product measure, 82, 418, 422, 415–422, 488, 488, 489, 498, 506, 531, 648, 704
of random variables, 231
space, 82, 83, 85, 87, 422, 486, 489, 648
- Projection into subspace of L^2 , 485
lemma, 285, 288
of random point on sphere, 285–287
of set into subspace, 626, 627
slice, 626
- Proper distribution, 541, 551, 553, 705, 717
- Public key cryptography, 584
- Pure birth process, 312
- Pythagoras’s theorem, 522
- \mathbb{Q} , *see* Rational numbers
- \mathbb{Q} -function, 316
- Quadrant, 607
- Quadratic form, 608
of a normal, 324, 365
- Quantile, 760, 761, 763, 766
- Quantum mechanics, 768
- Quarantining, 687
- Quaternary expansion, 640, 641
- Queueing process, 55, 543
theory, 298, 303
- Quicksort, 455–459
- Quotiented space, 484, 485
- $\bar{\mathbb{R}}$, *see* Extended real line
- \mathbb{R}^∞ , *see* Infinite product space
- ρ_1 , ρ_r , *see* Hamming distance
- ρ_0 , *see* Talagrand’s convex distance
- $R(\mathcal{J})$, *see* Ring generated by the half-closed intervals
- $\overline{R(\mathcal{J})}$, definition, 387
- $R(k, k)$, $R(j, k)$, *see* Ramsey number
- Radar maps of Venus, 337, 338, 340, 341
- Rademacher functions, 143, 141–146, 149–152, 160, 177, 178, 496, 594, 605, 786
are independent, 144–148
are orthogonal, 143, 154, 157, 159
- Rademacher, H. A.*, 140, 142, 594n
- Radio scintillation, 511
- Radon–Nikodym theorem, 464
- Ramanujan, S.*, 141, 582n
- Ramsey number, 109, 131, 136, 138
theory, 109, 130
- Ramsey’s theorem, 109
- Ramsey, F. P.*, 109
- Random arc, 283, 284
direction, *see* Random point on circle (sphere)
flight, 285, 284–288, 517, 752
geometric graph, 681, 681, 685
graph, 10, 85, 86, 119, 138, 394, 568–571, 650, 665
mating, 78
number generator, 761
pair, 200, 211
permutation, 66, 132, 597, 608, 668, 672, 674, 676, 678, 745, 747, 767
point in interval, 280, 281, 311, 453, 679, 682
point in unit square, 649
point on circle, 281, 311, 425, 434, 679, 686
point on sphere, 284, 284–288, 425, 752
process, 340
sample, 330, 603, 753
selection, 67
set, 31, 90, 91
variable, 197, 396, 393–426, 445
vector, 412, 414
walk, 183, 244, 244–252, 274, 285, 542, 550, 555, 550–557, 559, 562, 598, 690, 721, 751–753, 762, 763
- Randomisation, 221, 232, 313
- Rank (in cards), 10
- Rare event (Poisson approximation), xvi, 119, 570, 660
- Ratio of random variables, 232
- Rational atoms, 402
form, 560
- Rational numbers are countable, 27, 377, 400
are dense, 377, 596n
- Ray, 599, 600, 602, 608
- Rayleigh’s density, 320
- Rayleigh, J. W. S.*, 286
- Rayleigh–Chandrasekhar density, 287, 288
- Record values, 312, 767
- rect, *see* Rectangular function
- Rectangle in the plane, 212
 n -dimensional, 26, 217
- Rectangular density, 276
function, 161, 170, 178, 181, 182
- Recurrent event, 451n
- Red Auerbach, disdain for basketball statistics, 33n
- Reductionist theorem, 704, 723
- Refinement of a family of sets, 28
of a partition, 430
- Regression, 220
- Regular expressions, 274
- Relative entropy (Kullback–Leibler divergence), 481
- Reliable replication, 615
- Remote event, 422, 423
- Renewal epoch, 534, 559
equation, 532–536, 540, 553, 558
function, 535
process, 532–536, 536, 543, 544, 553, 559
- Rényi, A.*, 86n, 121n, 123
- Repeated independent trials, 82

- Reservoir sampling, 233
 Residual time, 298, 301, 546
 Returns to the origin, 245,
 247, 250, 252, 274, 555
 Reversed walk, 248, 554, 589
Rezucha, I., 233n
 Rice's density, 320, 363
Richter, W., 750n
Riemann, B., 579
 Riesz representation theorem, 708
Riesz, F., 483
 Right continuous function,
 definition, 778
 Ring, 381
 generated by the half-closed intervals, 15, 381,
 390, 416
 Robbins's bounds for the factorial, 575, 658, 673
Robbins, H. E., 311, 577n
Rodemich, E. R., 195n
Rosenblatt, M., 768n
 Rotation transform, 227, 329,
 332, 339, 342, 756, 757,
 768
 Roulette, 288
 Round number, 584
 Row sum, 674, 718, 737, 738,
 742, 744, 746, 747
 Royal flush (poker), 272
Ruciński, A., 688n
 Rumours (spread of), 32
 Run length, 262–264, 423, 426
 Runs, 90, 92, 559, 560, 767
Runyon, D., 585
Ryser, H. J., 135n, 135

Sagan, H., 641n
Salehi, A. T., xxiii
 Sample
 function, *see* Sample path
 mean, 241, 331, 334, 566,
 567
 median, 314
 path of Brownian motion,
 355, 361
 path of Poisson process,
 302
 path of random walk, 245,
 587
 point, 9
 size, 242, 567, 603, 608
 space, 9–12
 variance, 331, 333, 517
 Sampling
 with replacement, 79, 99
 without replacement, 84
Samuelson, P. A., 349
Savage, J., 349
 Scale
 factor, 284
 operator, 167
 parameter, 207
 property of transforms,
 167, 171, 172, 524, 530,
 726
Schell, E. D., 239n
 Schläfli's theorem, 607
Schläfli, L., 607n
Scholes, M. S., 349
Schwarz, H. A., 787n
 Second moment, 472
 Secretary problem, *see* Marriage problem
 Selection
 matrix, 674–678
 principle, *see* Helly's selection principle
 Self-adjoint, *see* Symmetric
 Self-similarity, 308
 Semigroup of operators, 703–704
 Semiring of sets, 34
 Sensor network, 681
 Sequential decision theory, 69
 Service rate, 546
 Set
 function, 16, 19, 23, 34,
 370–372, 374–376, 380,
 381, 384, 385, 389, 390,
 398, 399, 463, 468, 488,
 508, 599, 691, 692
 of measure zero, 156, 159,
 407, 374–718
 operations, 12–13
Shahrampour, S., xxiii
 Shannon's theorem, 506, 616
Shannon, C. E., 505n, 521n,
 612
Shapley, L. S., 69n
 Shattering a sample (or set),
 600, 601
 Shear transform, 227
 Sherrington–Kirkpatrick spin glass model, 650
Shevtsova, I. G., 730n
 Shot noise, 718
Shrikande, S. S., 135n
 Side information, *see* Conditional probability
Sierpiński, W., 347
 Sieve of Eratosthenes, 99–102
 Sifting theorem, 174
 σ-algebra
 definition, 15
 generated by a family of sets, 24, 397
 generated by a random variable, 397, 419
 σ-finite measure, 372, 374,
 391, 486–489
 Signal vector, 501–506, 519,
 647, 648
 Simple
 function, 474, 486, 487, 489
 random variable, 402, 429–433, 437, 438, 442, 464, 475
 Simpson's paradox, 40–42
 rule, 566
Simpson, E. H., 42n
 Simultaneous diagonalisation, 365
 sinc, *see* Sinc function
 Sinc function, 161, 170
 Single-server queue, 303
 Singular continuous distribution, 406, 434, 444, 508
Skitovich, V. P., 344n
 Skitovich–Darmois theorem, 344–346
 Slogans, 183, 302, 341, 429,
 455, 529, 536, 568, 603,
 614, 616, 617, 624, 651,
 656, 733
 Smallest σ-algebra, *see* σ-algebra generated by
 Smooth function, 692, 698,
 702, 724, 741
 Snedecor's density, 337
 Sojourn time, 304, 558
 Space
 exploration, why it would be wise, 64
 of continuous functions, 11
 of square-integrable functions, 11, 352, 353, 395,
 521, 594, 783
 Space-filling

- curve, 638, 641, 642, 649
 tour, 642–645
Spacings, 280–686
 small, 687
Span, 130, 198
Sparre Andersen, E., 307, 546, 555
Speckle interferometry, 511
Spectrum, *see Fourier transform*
Spencer, J., 112, 134n, 688n
Sphere-hardening, 503–505, 624
Spherical coordinates, 518
Spin
 interaction, 650
 quantum state, 650
Spitzer, F., 555
Spotlight theorem, 174, 175
Spreading sequence, 192
Square-root (of a positive definite matrix), 339
St. Petersburg game, 585–587
 with shares, 606
Stability under convolution of Cauchy, 309, 767
 of normal, 195, 318, 326, 723, 724
 of Poisson, 196, 256, 767
Stable distributions, 365, 724n
Standard
 deviation, 199, 208, 472
 normal density, 164, 171, 650
 representation of simple functions, 429
Star diameter, 511
Star Wars, *see Strategic Defense Initiative*
State (of chain, system), 52–55, 68
Stationary
 distribution, 53–55, 68, 80
 process, 341, 366, 473
Statistical tests for capture–recapture, 269, 274
 invasive species, 242
 periodogram analysis, 365
 polls, 240, 243
 population categories, 759–763, 765
 quality control, 269
sera and vaccines, 240
success runs (hot hand), 33, 560
Steele's theorem, 608
Steele, J. M., xxiv, 482, 485n, 608n
Steepest descent, 520
Stein's equation
 for the normal, 732, 733, 734
 for the Poisson, 656, 657, 659, 661, 669, 732
Stein's method, xvi, 730, 731, 751
Stein, C., xvi, 651n, 668, 730, 751n
Stein–Chen method, 652, 656–657, 662, 674, 677, 681, 685n, 731
Steinhaus's problem, 594
Steinhaus, H., 261, 593
Stieltjes integral, 468
Stirling numbers of the second kind, 137
Stirling's formula, 112, 153, 185, 195, 458, 496, 557, 575, 577, 768
Stochastic domination, 528
Stop-and-wait protocol, 258
Straight, straight flush (poker), 31
Strategic Defense Initiative, political whimsy, 645
Stratification, 313
Strong law of large numbers, 17, 158, 162, 363, 588, 595, 599, 605, 616, 635, 646
Student, *see Gosset, W. S.*
Student's density, 335, 365
Sub-Gaussian bound for the longest increasing subsequence, 638
Subadditivity
 of measure, 34, 355, 373
 of outer measure, 386
Subspace of vector space, 321, 757
Success
 in Bernoulli trial, 236
 run, 32, 33, 559, 560
Successive differences, 659, 660, 663
Successor (of vertex), 622, 623, 627n, 687
Sudoku, 134
Suit (in cards), 9
Sums of indicators, 660–688
Super-additivity, 446
Support of a function, 207
Supremum, 27, 773
 norm, *see Norm of function*
Surface area
 of sector of sphere, 285
 of sphere in n dimensions, 493
Surjection, 640
Swinburne, A., 359
Symmetric
 Bernoulli trial, 236, 562
 difference, 12
 distribution, 509, 550, 559
Symmetrisation
 by pairwise exchanges, 597, 608
 by permutation of double sample, 608
 inequality, 510, 597
Symmetrised distribution, 509–511, 597
T, definition, 281
t-statistic, 333–335, 337
Tail
 event, 153, 423
 probability, 153
 σ -algebra, 423
Talagrand's
 convex distance, 621, 623, 624, 630, 633, 635, 645, 649
 induction method, 624–633, 648
 theorem, 624, 630, 632, 635, 637, 645
Talagrand, M., xvii, 609, 617n, 619, 647
Tan, K., 254n
Tarry, G., 135
Tassa, T., 653n
Teicher, H., 768n
Temperature, 650
Tennis
 ranking, 264, 273
 tie-breaks, 273
Terminal ladder index, 547, 548, 549, 551, 552

- Terminating binary expansion, 141
 Ternary digit, 159
 Test
 for independence, 521
 statistic, 755
 Theorem of
 inclusion-exclusion, 94
 total probability, 42, 44, 47,
 48, 50, 53, 56, 57, 60, 65,
 67, 75, 82, 133, 221, 222,
 464, 511
 Third moment, 724, 725
Thoday, J. M., 137n
 Three series theorem, 595
 Three-of-a-kind (poker), 31
 Threshold function, 124, 267,
 570, 666, 681, 685, 687
Tian, L., 557n
 Toroidal metric, 683, 685
 Torus, 683, 684
 Total
 σ -algebra, 16, 24–26
 variation distance, 646,
 656, 662, 685, 718
 Totally ordered set, 455
 Totient function, 100, 101
 Tour (of cities), 641–643
 Tournament, 91
 Tower property, 464
 Trace, 599, 601
 Transition probability, 52, 53,
 55, 65, 66, 68, 70
 Transitive property of total
 order, 455
 Translate of a set, 392
 Translation
 invariance of Lebesgue
 measure, **292**, 294, 376,
 392, 504
 operator, 167
 property of transform, 167
 trap, *see* Trapezoidal function
 Trapezoidal function, 170,
 181, 182
 Travelling salesman prob-
 lem, 641–646, 649
 Tree, **103**, 122, 137
 Treize, 677, 685
 Triangle
 characteristic property, 277
 in random graph, 138, **568**,
 568–571, 665, 666, 688
 inequality, 362, 401, 406,
 417, 483, **484**, 510, **520**,
 586, 592, 593, 596, 598,
 603, 643, 663, 694, **697**,
 701, 702, 704, 707, 708,
 724, 728, 732, 735, 741,
 782, **785**, 787
 Triangular
 array, 718, **737**, 738, 742,
 746, 750
 density, 233, **278**, 510
 function, **170**, 726
 system of wavelets, 355,
 356, 366
 Trigonometric
 polynomial, 365
 series with random coeffi-
 cients, 351
 Trinomial distribution, **272**,
 556
 Trivial σ -algebra, 15
Trotter, H. F., 720n
 TSP, *see* Travelling salesman
 problem
Turán, P., 582, 583n
Tversky, A., 33n
 Two pair (poker), 31
 Two-colouring of the edges
 of a graph, 109
 Two-dimensional distribu-
 tion, 391, 412
 Type of distribution (den-
 sity), **207**, 209, 276, 287,
 308, 315, 725, 739
Ulam, S., 567
 Unbiased
 estimator, 241, 331, 333,
 517, **521**
 sample, 244
 Uncorrelated
 normals, 328, 329, 476
 random variables, **213**,
 334, **474**, 475, 477, 490,
 610, 726, 734
 Uncountable set, 377, 380
 Uniform
 atomic distribution, 718
 density, 24, 205, 234, **276**,
 403, 426, 510, 604, 721
 distribution, 214, 220, **276**,
 276–280, 285, 286, 290–
 292, 306, 313, **403**, 420–
 422, 433, 443, 461, 516,
 524, 525, 604, 647, 679,
 682, 718, 738, 739
 distribution in a cube, 395,
 682
 distribution in unit square,
 649
 distribution on circle, **281**,
 436, 714, 715, 717
 distribution on the sphere,
 768
 prior, 62, 64, 275, 306
 Uniformly continuous func-
 tion, 160, 354, **779**, 780,
 782
 Union
 bound, *see* Boole's inequal-
 ity
 of sets, 12
 Uniqueness of measure in-
 duced by a distribution,
 390
 Unit
 cube in n dimensions, 395,
 614, 635
 disc function, 515
 sphere, 646, 768
 step function, *see* Heaviside
 function
 University of Pennsylvania,
 ivy encrusted institution
 in the Colonies, Ben's
 brainchild, xvii, xxiv
 Unreliable transcription, 614,
 616
 Unusual dice, 557
 Upper derivative, 361
 Urn problem, 7, 10, 31, 59,
 66–68, 72, 92, 95, 119,
 232, 238, 424, 450, 453,
 462, 463
 indistinguishable balls, 31
Valiant, L., 604n
 Valley (energy minimum),
 518
Vallone, R., 33n
 Van der Waerden number,
 113, 130, 131
 Vandermonde's convolution,
 269, **271**, 510
Vandermonde, A., 271
Vapnik, V. N., xvi, 599n
 Vapnik-Chervonenkis
 class, 599, **600**, 601, 603,
 607, 608
 dimension, **601**, 602, 608,
 650

- theorem, 600, 602, 603, 608
 Variance
 definition, 199, 208, 212, 472
 of binomial, 239
 of exponential, 293
 of geometric, 258
 of negative binomial, 260
 of normal, 316
 of Poisson, 255
 of uniform, 276
 Vector
 inequality, 28, 216
 space of operators, 700–703
 sum of random directions, 285, 287, 436, 752
Venkatesh, A. B., *v*
Venkatesh, C. S., *v*
Venkatesh, S. J., *v*
Venkatesh, S. S., 193n, 195n, 233n, 254n, 518n, 685n, 751n
Venkatesh, V. M., *v*
 Venn diagram, 12, 75
 Venus, *see* Radar maps of Venus
Vergetis, E., xxiii
 Vertex of a graph, 85, 102, 394, 650
 Viète's formula, 140, 141, 144, 145–148, 150, 159, 594
Viète, F., 139–141
 Violins, why two are twice as loud as one, 288
Virgil, 498
 Vol, *see* Volume of a ball (set)
 Volume
 of a ball, 295, 492–494, 496, 498–500, 504, 519, 684
 of a parallelepiped, 481
 of a set, 160, 395, 482, 499, 500, 504, 505
von Bortkewitsch, L., 253
von Mises, R., 9n
von Neumann, J., 426n, 567
 $w(k; p)$, $w_t(k; p)$, *see* Waiting time distribution
 Waiting time, 205, 234, 262, 267, 273, 298, 303, 312, 451, 454, 510, 534, 543, 545, 721, 742
 distribution, 20, 256–262, 312, 427, 545
 Wall Street, predicting stock movements, 762
Wallis, J., 141
 Walsh–Kaczmarz function, 160
Warmuth, M. K., 604n
 Waveform with random phase, 473
 Wavelets, 352
 Weak law of large numbers, xvii, 80, 154, 155, 158, 160, 162, 195, 292, 409, 426, 462, 564, 585, 690, 767
 for stationary sequences, 605
 Weakest link of chains, 294
 Weierstrass's approximation theorem, 162, 572, 605, 710
Weierstrass, K., 346, 404, 571n, 773, 777
 Weight vector, 608
Weinberg, S., 31n
Wendel, J. G., 607n
 Weyl's equidistribution theorem, 196, 713, 714, 717
Weyl, H., 196, 713n
Whitworth, W. A., 46n, 247
 Wiener process, 352
Wiener, N., 348n, 351
Wilf, H. S., xxiv, 653n
 Wireless communication, 191
 Witness (for sequence), 636, 637
Wolf, R., 292
Wright, E. M., 581n, 715n
Wu, Z., xxiii
You, J., 557n
 z-statistic, 331
 Zero-one law, 423, 594
 Zipf distribution, 308
Zipf, G. K., 308
Zygmund, A., 351

