An Invitation to Modern Number Theory

Steven J. Miller and
Ramin Takloo-Bighash

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Contents

Preface

This book on modern number theory grew out of undergraduate research seminars taught at Princeton University (2001–2003), and similar courses taught at New York University, Ohio State, Brown University and a summer Research Experience for Undergraduates at the American Institute of Mathematics. The purpose of these classes was to expose undergraduates to current research topics in mathematics. To supplement the standard lecture-homework classes, we wanted a course where students could work on outstanding conjectures and open problems and experience firsthand the kinds of problems mathematicians study. In the sciences and engineering, undergraduates are often exposed to state of the art problems in experimental laboratories. We want to bring a similar experience to students interested in mathematics. This book is the outcome of that effort, providing the novice with hints as to what we feel is a good path through the immense landscape of number theory, as well as the needed background material. We have tried to give students and their teachers a model which can be used to develop their own research program; to this end, throughout the book are detailed descriptions of accessible open problems and references to the literature. Though we encourage students and teachers to attempt some of the open problems, the book stands alone and may be used for a standard lecture course (especially for new subjects such as Random Matrix Theory where there are not many introductory works accessible to undergraduates). Our goal is to supplement the classic texts in the field by showing the connections between seemingly diverse topics, as well as making some of the subjects more accessible to beginning students and whetting their appetite for continuing in mathematics.

The book has five parts, though several themes run throughout the book.

- Part I deals with basic number theory (cryptography and basic group theory), elementary L-functions (including the connections between zeros of $\zeta(s)$ and primes), and solutions to Diophantine equations. The material in this part is fairly standard, and could serve as an introduction to number theory. In some sections a little group theory and first semester complex analysis is assumed for some advanced topics. Our purpose in the first chapter is not to write a treatise on cryptography, but to review some of the background necessary from basic number theory for later chapters. It is possible to *motivate* this material in the context of cryptography; though these applications are very important, this connection is meant only to interest the reader, as this is not a exposition on cryptography. Similarly, elliptic curves are a terrific example for some of the material in Chapter 4 (and later in the book); as such, we introduce just enough for these purposes. As there are numerous excellent books on both of these subjects, we have kept our treatments short and refer the interested reader to these for more details. One theme in these chapters is the search for efficient algorithms, which appears frequently in later parts as well.

- Part II has two connected themes: approximating numbers with rationals, and continued fractions. In the first, the basic properties of algebraic and transcendental numbers are discussed, and a proof of Roth's Theorem (on how well algebraic numbers can be approximated by rationals) is given in full detail. This is one of the great achievement of 20th century number theory. Roth's Theorem has now been greatly generalized, and there are a few different ways to prove it. Our formulation and proof follow Roth's original proof. The proof we present here, though long and technical, requires only knowledge of elementary calculus and linear algebra. The second part is an introduction to continued fractions (a subject of interest in its own right, but also of use in approximation theory) and culminates in several open problems; this chapter is independent of Roth's Theorem and may serve as a survey to the subject. Also, time and again (especially in Part III when we study digit bias and spacings between terms in certain sequences), we see that answers to many number theoretic questions depend on properties of the numbers in the problem; often the continued fraction expansion highlights these properties. There are references to open problems in continued fractions, many of which concern the distribution of digits (see Part III).

- Part III encompasses three themes. The first is the distribution of the first digit of several interesting sequences (for example, the Fibonacci numbers and iterates of the $3x + 1$ map). We use this problem as a motivation for hypothesis testing (whether or not numerical data supports or contradicts conjectured behavior). Hypothesis testing is an extremely important subject, especially as computers are used more and more frequently in mathematics. The second theme centers around the Gauss-Kuzmin Theorem for the distribution of digits of continued fractions. We then develop enough Fourier Analysis to prove various basic results, including a sketch of the proof of the Central Limit Theorem and Poisson Summation (one of the most used tools in number theory). We use these results to investigate the behavior of $n^k\alpha \bmod 1$ for fixed k, α (specifically, the spacings between these numbers in $[0, 1]$; for many k and α these spacings appear to be the same as the spacings between adjacent primes); we study other spacing problems in Part V; in fact, our results on the Fourier transform are needed in Chapter **??** when we investigate zeros of L-functions. Numerous open problems and references to the current literature are provided.

- Part IV is a brief introduction to the Circle Method, a powerful theory to study questions in additive number theory (such as writing a number as a sum of a fixed number of k^{th} powers or primes). After developing the basics of the theory, we discuss in some detail why, using these methods, we cannot (yet?) show that any even number is the sum of two primes but we can show any large odd number is the sum of three primes. We use the Circle Method to predict how many Germain primes (p and $\frac{p-1}{2}$ both prime) are less than x. This example illustrates many of the key techniques of the theory, as well as the problems that arise in applications. Further, the density of these primes has recently been connected to fast primality testing algorithms. As usual we conclude with some open problems.

- Part V is an introduction to Random Matrix Theory and its interplay with number theory. What began as a model in the 1950s for physicists to study the energy levels of heavy nuclei has become a powerful tool after a chance encounter one day at tea in the 1970s (see [**?**] for an entertaining account of the meeting) for predicting the behavior of zeros of $\zeta(s)$ and other L-functions; knowledge of these zeros is intimately connected to properties of primes. The general result is that there is a striking similarity between the spacings between energy levels of heavy nuclei, eigenvalues of sets of matrices and zeros of L-functions. We take a classical approach to the subject. Results from linear algebra and occasionally first semester complex analysis are used (especially in the final chapter); a review of enough of the background material is provided for students to follow the key ideas in the proofs. There are numerous open problems requiring only elementary probability theory and linear algebra (at the level covered in this book); many have already been successfully investigated by our students.

There are several chapters throughout the book covering background material in basic number theory, algebra, Fourier analysis and probability theory, as well as two appendices on needed calculus, analysis and linear algebra results. Clearly our book is not meant to replace standard textbooks in these fields. We have two reasons for including these background chapters (in addition to the material being interesting in its own right). First, waiting for students to assemble such a background takes time, and the main purpose of our book is to show students in the early stages of their education what mathematicians do, and the interplay between the various parts of number theory and mathematics. Second, often very little of the background subjects is needed to understand the basic formulation and set-up of current work. Therefore a student who has not seen such material in a previous course can get a feel for these subjects by reading the review and background chapters, and then move on to the current research chapters. We have, however, written the chapters in such a way that there are often additional remarks or sections for students with stronger backgrounds. We have also included references throughout the book showing how the same methods and techniques are used for many different problems.

We have strived to keep the pre-requisites to a minimum: what is required is more a willingness to explore than a familiarity with the landscape. Several times we use results from later in the book in earlier investigations; our hope is that after seeing how these theorems are used and needed the reader will be motivated and interested enough to study the proofs. For most of the book one-variable calculus is the only requirement. We have also tried to emphasize common techniques in proofs (the reader is strongly encouraged to study the *techniques* entry in the index).

The book breaks naturally into five parts. Depending on the background of the students, and whether or not a class is going to explore open problems further, a typical semester class would cover material from one part of the book (as well as whatever background material is needed), though we recommend everyone at least skim Chapter **??** to ensure familiarity with the language and some of the motivating influences and themes of number theory. Many topics (such as applications to cryptography, algebraic structure of numbers and spacings between events) occur in various forms throughout the book. In a two semester course, one can cover two of the advanced parts and see these connections. We have also tried to give students the opportunity to discover the theory by themselves by giving many exercises. Mathematics is not meant to be a passive pursuit. Some of the problems are mere warm-ups; others are real problems that require time and effort. The reader should not be discouraged at being unable to work out all the problems. The value of an exercise is often in the time and energy spent on it, rather than the final solution. Many of the more difficult problems are standard theorems and can be seen proved in other textbooks. In this regard our manuscript is in the spirit of [Mu2]. **In Appendix B we have provided hints and further remarks to certain exercises; these problems are marked with either an (h) or (hr) in the text.**

We have assembled an extensive bibliography to aid the reader in further study. In addition to the excellent texts [AZ, Apo, BS, Da1, Da2, EE, Est2, Fe, HW, IR, IK, Kh, Kn, La2, Meh2, Na, NZM, ST, vdP6] on continued fractions, number theory and random matrix theory, we recommend the recent work of Narkiewicz [Nar] (where the reader will find proofs of many number theory results, as well as over 1800 references) as well as [Guy] (where there are extensive bibliographies for open problems). We conclude in Appendix C with some remarks on common themes running through this book and number theory.

The students in our courses used computers to assemble large amounts of data for some of the problems mentioned in the text, which then led us to appropriate conjectures and in some cases even gave us ideas on how to prove them. For links to previous student reports as well as some of the research papers mentioned in the bibliography, please visit

http://www.math.princeton.edu/mathlab/book/index.html

These include student programs (mostly in C++, Maple, Mathematica, MATLAB, or PARI) and detailed references for those interested in continuing these studies. Students should also consult MathSciNet [AMS], the arXiv [Cor1] and Project Euclid [Cor2] to find and download additional references.

It is a pleasure to thank the professors and teaching assistants who have helped run the class over the years (Alex Barnett, Vitaly Bergelson, João Boavida, Alexander Bufetov, Salman Butt, Brian Conrey, David Farmer, Harald Helfgott, Chris Hughes, James Mailhot, Atul Pokharel, Michael Rubinstein, Peter Sarnak, Lior Silberman, Yakov Sinai, Warren Sinnott, Florin Spinu and Andrew Wiles), as well as the students.

We would also like to thank several of our colleagues. In particular, we thank Eduardo Dueñez, Rob Gross and Amir Jafari for reviewing an early draft and providing numerous helpful suggestions to improve the presentation, and Timothy Abbot, Mike Buchanan, Scott Craver, Kevin Dayaratna, Dean Eiger, Manfred Einsiedler, Dan File, Chris Hammond, Ted Hill, Alex Kontorovich, Josh Krantz, Matt Michelini, Jeff Miller, Liz Miller, Paria Mirmonsef, C. J. Mozzochi, Anna Pierrehumbert, Amitabha Roy, Zeév Rudnick, Eitan Sayag, Aaron Silberstein, Dan Stone, Howard Straubing, Yuri Tschinkel, Akshay Venkatesh and Bill Zaboski for discussions and comments on various chapters. The first author gave several lectures on the material to a summer research group and the Ross Program at Ohio State (summer 2004), and is indebted to the students for their comments and suggestions. We are grateful to Nicole, Michelle and Leo Beaupre, Andrew and David Norris, Joe Silverman and the staff at Princeton University Press for help with the illustrations, and to Stephen Kudla for mutually productive LaTeX discussions.

We are extremely grateful to Princeton University Press, especially to our editor Vickie Kearn, our production editor Lucy Day W. Hobor and our copyeditor Jennifer Slater, for all their help and aid, to Bob Gunning, for initiating contact between us and PUP and encouraging us to write the book, and to the National Science Foundation's VIGRE program, which helped fund many of the classes at Princeton, NYU and Ohio State.

The first author was partially supported by VIGRE post-doctoral fellowships at Princeton, New York University, The Ohio State University and Brown University, and enjoyed the hospitality of Boston University during the final stages of the project. The second author enjoyed the hospitality of the University of Maryland at College Park, Johns Hopkins University and The Ohio State University at various stages of working on the project. His work was partially supported by a Young Investigator's Award from the National Security Agency.

Steven J. Miller
Providence, RI
December 2005

Ramin Takloo-Bighash
Princeton, NJ
December 2005

Notation

$\mathbb{W}$: the set of whole numbers: $\{1, 2, 3, 4, \dots\}$.

$\mathbb{N}$: the set of natural numbers: $\{0, 1, 2, 3, \dots\}$.

$\mathbb{Z}$: the set of integers: $\{\dots, -2, -1, 0, 1, 2, \dots\}$.

$\mathbb{Q}$: the set of rational numbers: $\{x : x = \frac{p}{q}, p, q \in \mathbb{Z}, q \neq 0\}$.

$\mathbb{R}$: the set of real numbers.

$\mathbb{C}$: the set of complex numbers: $\{z : z = x + iy,\ x, y \in \mathbb{R}\}$.

$\Re z$, $\Im z$: the real and imaginary parts of $z \in \mathbb{C}$; if $z = x + iy$, $\Re z = x$ and $\Im z = y$.

$\mathbb{Z}/n\mathbb{Z}$: the additive group of integers mod n: $\{0, 1, \dots, n-1\}$.

$(\mathbb{Z}/n\mathbb{Z})^*$: the multiplicative group of invertible elements mod n.

$\mathbb{F}_p$: the finite field with p elements: $\{0, 1, \dots, p-1\}$.

$a|b$: a divides b.

$p^k||b$: p^k divides b and p^{k+1} does not divide b.

(a, b) : greatest common divisor (gcd) of a and b, also written $\gcd(a, b)$.

prime, composite : a positive integer a is prime if $a > 1$ and the only divisors of a are 1 and a; if $a > 1$ is not prime, we say a is composite.

coprime (relatively prime) : a and b are coprime (or relatively prime) if their greatest common divisor is 1.

$x \equiv y \mod n$: there exists an integer a such that $x = y + an$.

$\forall$: for all.

$\exists$: there exists.

Big-Oh notation : $A(x) = O(B(x))$, read “$A(x)$ is of order (or big-Oh) $B(x)$”, means $\exists C > 0$ and an x_0 such that $\forall x \geq x_0$, $|A(x)| \leq C\,B(x)$. This is also written $A(x) \ll B(x)$ or $B(x) \gg A(x)$.

Little-Oh notation : $A(x) = o(B(x))$, read “$A(x)$ is little-Oh of $B(x)$”, means $\lim_{x\to\infty} A(x)/B(x) = 0$.

$|S|$ or $\#S$: number of elements in the set S.

p : usually a prime number.

i, j, k, m, n : usually an integer.

$[x]$ or $\lfloor x \rfloor$: the greatest integer less than or equal to x, read “the floor of x”.

$\{x\}$: the fractional part of x; note $x = [x] + \{x\}$.

supremum : given a sequence $\{x_n\}_{n=1}^{\infty}$, the supremum of the set, denoted $\sup_n x_n$, is the smallest number c (if one exists) such that $x_n \leq c$ for all n, and for any $\epsilon > 0$ there is some n_0 such that $x_{n_0} > c - \epsilon$. If the sequence has finitely many terms, the supremum is the same as the maximum value.

infimum : notation as above, the infimum of a set, denoted $\inf_n x_n$, is the largest number c (if one exists) such that $x_n \geq c$ for all n, and for any $\epsilon > 0$ there is some n_0 such that $x_{n_0} < c + \epsilon$. If the sequence has finitely many terms, the infimum is the same as the minimum value.

$\Box$: indicates the end of a proof.

PART 1
Probabilistic Methods and Equidistribution

Chapter One

Introduction to Probability

In this chapter we give a quick introduction to the basic elements of Probability Theory, which we use to describe the limiting behavior of many different systems; for more details see [Du, Fe, Kel]. Consider all numbers in $[0,1]$. Let $p_{10,n}(k)$ be the probability that the n^{th} decimal (base 10) digit is k for $k \in \{0,\dots,9\}$. It is natural to expect that each digit is equally likely. This leads us to conjecture that $p_{10,n}(k) = \frac{1}{10}$ for all n. There is nothing special about base 10 — the universe does not care that we have ten fingers on our hands. Thus if we were to write our numbers in base b, then $k \in \{0,1,\dots,b-1\}$ and it is natural to conjecture that $p_{b,n}(k) = \frac{1}{b}$. These statements can be easily proved. If we look at the n^{th} digit of 10 million randomly chosen numbers, we expect to see about 1 million ones, 1 million twos, and so on; we will, of course, have to specify what we mean by randomly. What about the fluctuations about the expected values? Would we be surprised if we see $1,000,053$ ones? If we see $1,093,127$? The answer is given by the Central Limit Theorem, stated in §1.4 and proved in §**??**.

Instead of choosing numbers randomly in $[0,1]$, what if we consider special sequences? For example, how is the *first* digit of 2^n base 10 distributed? The possible digit values are $1,\dots,9$. Are all numbers equally likely to be the first digit of 2^n? We see in Chapter 2 that the answer is a resounding no. Another possible experiment is to investigate the n^{th} decimal digit of $\sqrt{p}$ as p varies through the primes. Do we expect as $n \to \infty$ that each number 0 through 9 occurs equally often? Do numerical experiments support our conjecture? Building on this chapter, in Chapter 2 we discuss how to analyze such data.

The probability of observing a digit depends on the base we use. What if we instead write the continued fraction expansion (see Chapter **??**) of numbers in $[0,1]$? The advantage of this expansion is that it does not depend on a base *as there is no base!* What is the probability that the n^{th} digit of the continued fraction expansion equals k, $k \in \{1,2,\dots\}$? How likely is it that the n^{th} digit is large, say more than a million? Small? We can already answer this question for certain numbers α. If α is rational then it has a finite continued fraction expansion; if α is a quadratic irrational, it has a periodic expansion. What is true about the expansions of the other $\alpha \in (0,1)$? We answer such questions in Chapter **??**.

Let $\{x\}$ denote the fractional part of x. Thus $\{x\} = x \bmod 1$. Consider an irrational number $\alpha \in (0,1)$. For each N look at the N numbers $\{1\alpha\}, \{2\alpha\}, \dots,$ $\{N\alpha\}$. Rearrange the above $\{n\alpha\}$ in increasing order, and for definiteness label them $\beta_1,\dots,\beta_N$:

$$0 \ \le\ \beta_1 \ \le\ \beta_2 \ \le\ \cdots\ \le\ \beta_N. \tag{1.1}$$

As we have N numbers in $[0,1]$, the average distance between numbers is about

$\frac{1}{N}$. What does the spacing between adjacent β_i's look like? How often are two adjacent β_i's twice the average spacing apart? Half the average spacing apart? We prove some results and describe open problems in Chapter **??**, and then in Part **??** we investigate the spacings between eigenvalues of matrices, energy levels of heavy nuclei like Uranium and zeros of L-functions, showing connections between these very different systems!

1.1 PROBABILITIES OF DISCRETE EVENTS

We begin by studying the probabilities of discrete sets; for example, subsets of the integers or rationals or any finite set. Many interesting systems are discrete. One common example is flipping a coin a finite number of times; in this case we are often interested in the number of heads or tails. Another is to have time discrete; for example, people waiting in line at a bank, and every minute there is a chance a teller will serve the next person in line.

In the last example, if instead of measuring time in minutes we measured time in seconds or tenths of a second, for all practical purposes we would have a continuous process. While discrete sets are often good approximations to continuous processes, sometimes we actually need the continuous case; we describe continuous probability distributions in §1.2.3. We assume the reader is familiar with elementary set operations and countable sets (see §**??**).

1.1.1 Introduction

Definition 1.1.1 (Outcome Space, Outcomes)**.** *Let $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$ be an at most countable set. We call Ω the sample (or outcome) space, and the elements $\omega \in \Omega$ the outcomes.*

Thus, the outcome space is the collection of possible outcomes.

Example 1.1.2. *Flip a coin 3 times. The possible outcomes are*

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}. \tag{1.2}$$

If we flip a coin three times, how many heads do we expect to see? What is the probability we observe exactly three heads? Exactly two heads? The answer depends on the coin. If the coin is fair, for each flip we have a 50% chance of getting a head and a 50% chance of getting a tail. The coin, however, need not be fair. It could have some probability p of landing on heads, and then probability $1-p$ of landing on tails. For many investigations, we need more than just a collection of possible outcomes: we need to know how likely each possible outcome is.

Definition 1.1.3 (Probability Function)**.** *We say $p(\omega)$ is a **(discrete)** probability function or distribution on Ω if*

1. $0 \le p(\omega_i) \le 1$ *for all* $\omega_i \in \Omega$.
2. $\sum_i p(\omega_i) = 1$.

The first statement says that each outcome has a non-negative probability of occurring, and nothing can have a probability greater than 1 (a probability of 1 of happening means the event happens); the second statement quantifies the observation that something definitely happens.

We call $p(\omega)$ the probability of the outcome ω. Given an outcome space with a probability function, we can investigate functions of the outcomes.

Definition 1.1.4 (Random Variable)**.** *Let X be a function from Ω to $\mathbb{R}$. That is, for each outcome $\omega \in \Omega$ we attach a real number $X(\omega)$. We call X a random variable.*

A random variable is essentially a function of the outcomes, assigning a number to each outcome. As there are many functions that could convert outcomes to numbers, for any outcome space there are many random variables. With the same outcome space from Example 1.1.2, one possible random variable is $X(\omega)$ equals the number of heads in ω. Thus, $X(HHT) = 2$ and $X(TTT) = 0$. Additionally, for $i \in \{1, 2, 3\}$ let

$$X_i(\omega) \ = \ \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ toss is a head} \\ 0 & \text{if the } i^{\text{th}} \text{ toss is a tail.} \end{cases} \tag{1.3}$$

Note that

$$X(\omega) \ = \ X_1(\omega) + X_2(\omega) + X_3(\omega). \tag{1.4}$$

Remark 1.1.5 (Important)**.** The following situation occurs frequently. Consider the case when $\Omega \subset \mathbb{R}$ and X is a random variable. We often adjust our notation and write x for $\omega \in \Omega$; thus a capital letter denotes a random variable and a lowercase letter denotes a value it attains. For example, consider a roll of a fair die. The outcome space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the probability of each $\omega \in \Omega$ is $\frac{1}{6}$. Let X be the number rolled on the die. Then $X(1) = 1$, $X(2) = 2$, and so on. In this example, it is very convenient to call the outcome space the number rolled. The outcomes are the numbers 1, 2 and so on, rather then “the dice is a 1,” “the dice is a 2”; X is the random variable that is the number rolled, taking on values $x \in \{1, \dots, 6\}$. We shall mostly use $X : \Omega \to \mathbb{R}$ to represent a random variable and emphasize that the outcome space need not be a subset of $\mathbb{R}$, though the reader should be aware of both notations.

Example 1.1.6 (Important)**.** *Given an outcome space Ω with events ω with probability function p, p is a random variable.*

The terminology can be confusing, as a given random variable X is clearly not random — it is what it is! The point is we can attach many different random variable to a given Ω.

1.1.2 Events

Definition 1.1.7 (Events)**.** *We call a subset $A \subset \Omega$ an event, and we write*

$$\text{Prob}(A) \ = \ \sum_{\omega \in A} p(\omega). \tag{1.5}$$

Note each outcome is also an event.

Definition 1.1.8 (Range of X)**.** *The range of a random variable X is the set of values it attains, denoted $X(\Omega)$:*

$$X(\Omega) \ = \ \{r \in \mathbb{R} : \exists \omega \in \Omega \textit{ with } X(\omega) = r\}. \tag{1.6}$$

Note $X(\Omega)$ is the set of values attained by $X(\omega)$ as we vary $\omega \in \Omega$. Given a set $S \subset X(\Omega)$, we let $X^{-1}(S) = \{\omega \in \Omega : X(\omega) \in S\}$. This is the set of all outcomes where the random variable assigns a number in S.

Exercise 1.1.9. *Let Ω be the space of all tosses of a fair coin where all but the last toss are tails, and the last is a head. Thus $\Omega = \{H, TH, TTH, TTTH, \dots\}$. One possible random variable is X equals the number of tails; another is Y equals the number of the flip which is a head. Calculate the probabilities of the following outcomes in Ω. What is the probability that $X(\omega) \leq 3$? What is the probability that $Y(\omega) > 3$? What events do these correspond to?*

In general, we can associate events to any random variable. Let Ω be an outcome space with outcomes ω, and let X be a random variable. As we are assuming Ω is countable, the random variable X takes on at most countably many distinct values, so the range $X(\Omega)$ is at most countable. Let x_i denote a typical value. For each x_i, we can form the event $X(\omega) = x_i$; let us denote this event by A_i:

$$A_i \ = \ \{\omega \in \Omega : X(\omega) = x_i\} \ \subset \ \Omega. \tag{1.7}$$

Note that the A_i's are disjoint sets; if $\omega \in A_i \cap A_j$, then $X(\omega) = x_i$ as well as x_j. Further, $\cup_i A_i = \Omega$, because given any $\omega \in \Omega$, $X(\omega) = x_i$ for some i, hence ω is in some set A_i. The sets A_i form a **partition** of Ω (every $\omega \in \Omega$ is in one and only one A_i).

Remark 1.1.10 (Important)**.** By the above, given an outcome space Ω with outcomes ω and a probability function p and a random variable X, we can form a new outcome space $\widetilde{\Omega}$ with outcomes x_i with probability function $\widetilde{p}$ given by

$$\widetilde{p}(x_i) \ = \ \sum_{\substack{\omega \in \Omega \\ X(\omega) = x_i}} p(\omega). \tag{1.8}$$

Remark 1.1.11 (Important)**.** In a convenient abuse of notation, we often write

$$p(x_i) \ = \ p(X(\omega) = x_i) \ = \ \text{Prob}(\omega \in \Omega : X(\omega) = x_i). \tag{1.9}$$

We also call the random variable X an event, as the subsets of Ω corresponding to different values of X are events. Thus we can talk about the event "the value of the first roll," as the following example and Example 1.1.14 illustrate.

Example 1.1.12. *Let Ω be the set of all possible pairs of rolls of a fair die, and $X(\omega)$ equals the number of the first roll. We obtain events $A_1, \dots, A_6$. Let $Y(\omega)$ equal the number of the second roll, giving events $B_1, \dots, B_6$. If we consider the sum rolled, we have events $C_2, \dots, C_{12}$. For example, $C_7 = \{(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$, $(6,1)\}$. See Chapter 9 of [Sc] for a plethora of interesting problems on dice.*

Exercise 1.1.13. *Calculate the probabilities of the events $C_2, \dots, C_{12}$ for Example 1.1.12.*

Example 1.1.14 (Characteristic or Indicator Functions)**.** *We continue to reconcile our two notions of an event, namely a subset $A \subset \Omega$ and a random variable X. To any $A \subset \Omega$ we can associate a* ***characteristic*** *or* ***indicator random variable*** *1_A as follows:*

$$1_A(\omega) \ = \ \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{1.10}$$

Thus A is the set of ω where $1_A(\omega) = 1$.

Definition 1.1.15 (Complements)**.** *The complement of a set $A \subset \Omega$ is the set of all $\omega \notin A$. We denote this by A^c:*

$$A^c \ = \ \{\omega : \omega \in \Omega, \omega \notin A\}. \tag{1.11}$$

Using complements, we can rewrite the definition of the indicator random variable X_A:

$$X_A(\omega) \ = \ \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c. \end{cases} \tag{1.12}$$

Lemma 1.1.16. *Consider an outcome space Ω with outcomes ω and probability function p. Let $A \subset \Omega$ be an event. Then*

$$p(A) \ = \ 1 - p(A^c). \tag{1.13}$$

This simple observation is extremely useful for calculating many probabilities, as sometimes $p(A^c)$ is significantly easier to determine.

Exercise 1.1.17. *Prove Lemma 1.1.16. Consider 100 tosses of a fair coin. What is the probability that at least three tosses are heads?*

Exercise$^{\textbf{(hr)}}$ 1.1.18. *Consider 100 tosses of a fair coin. What is the probability that at least three consecutive tosses are heads? What about at least five consecutive tosses?*

Given an outcome space Ω with outcomes ω and random variable X, we can define a new random variable $Y = aX$, $a \in \mathbb{R}$, by $Y(\omega) = a \cdot X(\omega)$. This implies $p(Y(\omega) = ax_i) = p(X(\omega) = x_i)$. Thus if $X(\omega)$ takes on the values x_i with probabilities $p(x_i)$, $Y(\omega) = a \cdot X(\omega)$ takes on the values ax_i with probabilities $p(x_i)$.

Exercise 1.1.19. *Let X be a random variable on an outcome space Ω with probability function p. Fix a constant a and let $Y(\omega) = X(\omega) + a$. Determine the probability $Y(\omega) = y_i$.*

Example 1.1.20 (Geometric Series Formula)**.** *Alan and Barbara take turns shooting a basketball; first one to make a basket wins. Assume every time Alan shoots* *he has a probability $p \in [0,1]$ of making a basket, and each time Barbara shoots she has a probability $q \in [0,1]$ of making a basket. For notational convenience let $r = (1-p)(1-q)$. We assume that at least one of p and q is positive (as otherwise the game never ends); thus $r \in [0,1)$. The probability that Alan wins on his first shot is p, that he wins on his second shot is rp (he must miss his first shot, Barbara must miss her first shot, and then he must make his second shot), and in general that he wins on his n^{th} shot is $r^{n-1}p$. Letting x equal the probability that Alan wins, we find*

$$x \;=\; p + rp + r^2p + \cdots \;=\; p\sum_{n=0}^{\infty} r^n. \tag{1.14}$$

However, we also know that

$$x \;=\; p + (1-p)(1-q)x \;=\; p + rx. \tag{1.15}$$

This follows from observing that, once Alan and Barbara miss their first shots, it is as if we started the game all over; thus the probability that Alan wins after they each miss their first shot is the same as the probability that Alan wins (we must remember to add on the probability that Alan wins on his first shot, which is p). Since $x = p + rx$ we find $x = p/(1-r)$, so (1.14) *becomes*

$$\sum_{n=0}^{\infty} r^n \;=\; \frac{1}{1-r}, \tag{1.16}$$

the geometric series formula!

Exercise$^{\text{(h)}}$ 1.1.21. *The above example provides a proof for the geometric series formula, but only if $r \in [0,1)$. If $r < 0$ show how we may deduce the geometric series formula from the $r \geq 0$ case.*

Exercise$^{\text{(h)}}$ 1.1.22 (Gambler's ruin)**.** *Alan and Barbara now play the following game. Alan starts with n dollars and Barbara with m dollars (n and m are positive integers). They flip a fair coin and every time they get heads Barbara pays Alan a dollar, while every time they get a tail Alan pays Barbara a dollar. They continue playing this game until one of them has all the money. Prove the following:*

1. *If $n = m$ then the probability that Alan wins is $n/(n+m) = 1/2$.*
2. *If $n + m = 2^k$ for some positive k then the probability that Alan wins is $n/(n+m)$.*
3. *If $m = 2$ then the probability that Alan wins is $n/(n+m)$, and if $m = 1$ then the probability that Alan wins is $n/(n+m)$.*
4. *For $1 \leq m, n$ the probability that Alan wins is $n/(n+m)$.*

Investigate what happens for small m and n if the coin is not *fair.*

Remark 1.1.23. Exercises 1.1.20 and 1.1.22 provide examples of a useful technique, namely finding a relation for a probability p of the form $p = a + bp$ with a and b known.

Exercise$^{\text{(hr)}}$ 1.1.24. *Consider a circle of unit radius and a square of diameter 2. Assume we paint p percent of the perimeter blue and $1 - p$ of the perimeter red. Prove that if $p < 1/4$ then there* must *be a way to position the square inside the circle so that the four vertices are on the perimeter and all four vertices are on the red parts of the circle. Generalize the problem to an n dimensions.*

1.1.3 Conditional Probabilities

Consider two probability spaces Ω_1 and Ω_2 with outcomes ω_1 and ω_2. We can define a new outcome space

$$\Omega \ = \ \{\omega = (\omega_1, \omega_2) : \omega_1 \in \Omega_1 \text{ and } \omega_2 \in \Omega_2\}, \tag{1.17}$$

with outcomes $\omega = (\omega_1, \omega_2)$. We need to define a probability function $p(\omega)$, i.e., we need to assign probabilities to these outcomes. One natural way is as follows: let p_i be the probability function for outcomes $\omega_i \in \Omega_i$. We define

$$p(\omega) \ = \ p_1(\omega_1) \cdot p_2(\omega_2) \ \text{ if } \omega = (\omega_1, \omega_2). \tag{1.18}$$

Exercise 1.1.25. *Show the above defines a probability function.*

Of course, we could also define a probability function $p : \Omega \to \mathbb{R}$ directly. We again consider two tosses of a fair coin. We have outcomes $\omega = (\omega_1, \omega_2)$. Let us define $p(\omega) = \frac{1}{36}$, i.e., each of the 36 outcomes is equally likely. Let $X(\omega) = \omega_1$, the roll of the first die; similarly, set $Y(\omega) = \omega_2$, the roll of the second die.

Example 1.1.26. *What is* $\text{Prob}(X(\omega) = 2)$*? There are* 6 *pairs with first roll* 2*:* $(2,1), (2,2), \dots, (2,6)$*. Each pair has probability* $\frac{1}{36}$*. Thus,* $\text{Prob}(X(\omega) = 2) = \frac{6}{36} = \frac{1}{6}$.

More generally we have

$$\text{Prob}\left(X(\omega) = x_i\right) \ = \ \sum_{\substack{\omega = (\omega_1, \omega_2) \\ X(\omega) = x_i}} p\left(\omega\right). \tag{1.19}$$

The above is a simple recipe to find $\text{Prob}\left(X(\omega) = a\right)$: it is the probability of all pairs (ω_1, ω_2) such that $X(\omega) = x_i$, ω_2 arbitrary.

Let us consider a third random variable, the sum of the two rolls. Thus let $Z(\omega) = \omega_1 + \omega_2$, each outcome $\omega = (\omega_1, \omega_2)$ occurs with probability $\frac{1}{36}$. We have just seen that, if we have no information about the second roll, the probability that the first roll is a 2 is $\frac{1}{6}$ (what we would expect). What if, however, we know the sum of the two rolls is 2, or 7 or 10? Now what is the probability that the first roll is a 2? We are looking for pairs (ω_1, ω_2) such that $\omega_1 = 2$ and $\omega_1 + \omega_2 = 2$, 7, or 10. A quick inspection shows there are no pairs with sum 2 or 10. For a sum of 7, only one pair works: $(2, 5)$.

This leads us to the concept of **conditional probability**: *what is the probability of an event A, given an event B has occurred?* For an event A we can write

$$\text{Prob}(A) \ = \ \frac{\sum_{\omega \in A} p(w)}{\sum_{\omega \in \Omega} p(\omega)}. \tag{1.20}$$

Note the denominator is 1. For conditional probabilities, we restrict to $\omega \in B$. Thus, we have

$$\mathrm{Prob}(A|B) \;=\; \frac{\sum_{\substack{\omega\in A\\ \omega\in B}} p(w)}{\sum_{\omega\in B} p(\omega)}. \tag{1.21}$$

The numerator above may be regarded as the event $A \cap B$ (as both must happen, ω must be in A and B). $\mathrm{Prob}(A|B)$ is read *the probability of A, given B occurs* (or as the conditional probability of A given B). Thus,

Lemma 1.1.27. *If* $\mathrm{Prob}(B) \neq 0$,

$$\mathrm{Prob}(A|B) \;=\; \frac{\mathrm{Prob}(A\cap B)}{\mathrm{Prob}(B)}. \tag{1.22}$$

In the example above, let A be the event that the first roll is a 2 and B the event that the sum of the rolls is 7. As the die are fair, the probability of any pair (ω_1, ω_2) is $\frac{1}{36}$. Then

$$\begin{aligned} A &= \{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\}\\ B &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}\\ A\cap B &= \{(2,5)\}\\ \mathrm{Prob}(A|B) &= \frac{\mathrm{Prob}(A\cap B)}{\mathrm{Prob}(B)} \;=\; \frac{\frac{1}{36}}{6\cdot\frac{1}{36}} \;=\; \frac{1}{6}. \end{aligned} \tag{1.23}$$

Exercise 1.1.28. *Let Ω be the results of two rolls of two dice, where ω_1 is the number rolled first and ω_2 the number rolled second. For $\omega = (\omega_1, \omega_2) \in \Omega$, define the probabilities of the outcomes by*

$$p(\omega) \;=\; \begin{cases} \frac{1.5}{36} & \text{if } \omega_1 \text{ is even}\\ \frac{.5}{36} & \text{if } \omega_1 \text{ is odd.} \end{cases} \tag{1.24}$$

Show the above is a probability function of Ω. Let $X(\omega)$ be the number of the first roll, $Y(\omega)$ the number of the second roll. For each $k \in \{1, \dots, 6\}$, what is the probability that $Y(\omega) = k$ given $X(\omega) = 2$? Given $X(\omega) = 1$?

Exercise 1.1.29. *Three players enter a room and a red or blue hat is placed on each person's head. The color of each hat is determined by a coin toss, with the outcome of one coin toss having no effect on the others. Each person can see the other players' hats but not their own. No communication of any sort is allowed, except for an initial strategy session before the game begins. Once they have had a chance to look at the other hats, the players must simultaneously guess the color of their own hats or pass. The group shares a $3 million prize if at least one player guesses correctly and no players guess incorrectly. One can easily find a strategy which gives them a 50% chance of winning; using conditional probability find one where they win 75% of the time! More generally find a strategy for a group of n players that maximizes their chances of winning. See [Ber, LS] for more details, as well as [CS, LS] for applications to error correcting codes.*

1.1.4 Independent Events

The concept of **independence** is one of the most important in probability. Simply put, two events are independent if knowledge of one gives no information about the other. Explicitly, the probability of A occurring given that B has occurred is the same as if we knew nothing about whether or not B occurred:

$$\text{Prob}(A|B) \;=\; \frac{\text{Prob}(A \cap B)}{\text{Prob}(B)} \;=\; \text{Prob}(A). \tag{1.25}$$

Knowing event B occurred gives no additional information on the probability that event A occurred.

Again, consider two rolls of a fair dice with outcome space Ω consisting of pairs of rolls $\omega = (\omega_1, \omega_2)$. Let $X(\omega) = \omega_1$ (the result of the first roll), $Y(\omega) = \omega_2$ (the result of the second roll) and $Z(\omega) = X(\omega) + Y(\omega) = \omega_1 + \omega_2$ (the sum of the two rolls). Let A be the event that the first roll is 2 and B the event that the sum of the two rolls is 7. We have shown

$$\text{Prob}(A|B) \;=\; \frac{1}{6} \;=\; \text{Prob}(A); \tag{1.26}$$

thus, A and B are independent events. If, however, we had taken B to be the event that the sum of the two rolls is 2 (or 10), we would have found

$$\text{Prob}(A|B) \;=\; 0 \;\neq\; \text{Prob}(A); \tag{1.27}$$

in this case, the two events are not independent.

We rewrite the definition of independence in a more useful manner. Since for two independent events A and B,

$$\text{Prob}(A|B) \;=\; \frac{\text{Prob}(A \cap B)}{\text{Prob}(B)} \;=\; \text{Prob}(A), \tag{1.28}$$

we have

$$\text{Prob}(A \cap B) \;=\; \text{Prob}(A)\text{Prob}(B). \tag{1.29}$$

Note the more symmetric form of the above. In general, events $A_1, \dots, A_n$ are independent if for any subset $\{i_1, \dots, i_k\}$ of $\{1, \dots, n\}$ we have

$$\text{Prob}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \;=\; \text{Prob}(A_{i_1})\text{Prob}(A_{i_2}) \cdots \text{Prob}(A_{i_l}). \tag{1.30}$$

If events $A_1, \dots, A_n$ are pairwise independent, it is possible that the events are not independent.

Exercise 1.1.30. *Consider two tosses of a fair coin, each pair occurs with probability $\frac{1}{4}$. Let A be the event that the first toss is a head, B the event that the second toss is a tail and C the event that the sum of the number of heads is odd. Prove the events are pairwise independent, but not independent.*

Example 1.1.31. *Consider a fair die. Let A be the event that the first roll equals a, B be the event that the second roll equals b and C be the event that the sum of the two rolls is c, $c \in \{2, \dots, 12\}$. As each pair of rolls is equally likely, the probability that the first roll is a is $\frac{1}{6}$ (as six of the thirty-six pairs give a first roll of a). Thus,* *for any choices of a and b, the result of the first roll is independent of the second roll. We say that the two rolls (or the events A and B) are independent.*

Consider now event C, the sum of the two rolls. If the sum of the rolls is 7, then the probability that the first roll equals a is $\frac{1}{6}$ for all a; however, in general the conditional probabilities for the first roll will *depend on the sum. For example, if the sum is 2 then the probability that the first roll is 1 is 1 and the probability that the first roll is 2 or more is 0. Thus, events A and C (the first roll and the sum of the rolls) are not independent.*

Definition 1.1.32 (Independent Random Variables)**.** *Let X and Y be two random variables. We can associate events $A_i = \{\omega \in \Omega : X(\omega) = x_i\}$ and $B_j = \{\omega \in \Omega : Y(\omega) = y_j\}$. If for all i and j the events A_i and B_j are independent, we say the random variables X and Y are independent:* knowledge of the value of Y yields no information about the value of X.

Exercise 1.1.33. *Again consider two tosses of a fair coin, with $X(\omega)$ the number of the first toss and $Y(\omega)$ the number of the second toss. Prove X and Y are independent. Let Z be the random variable which is the number of heads in two tosses. Prove X and Z are not independent.*

The above exercise appears throughout probability investigations. For example, if we choose a non-rational $\alpha \in (0,1)$ "at random," we could let $X(\alpha)$ denote the value of the first decimal digit, and $Y(\alpha)$ denote the value of the second decimal digit. Are X and Y independent? The answer will depend on how we "randomly" choose α.

We give an example typical of the independence we will see in our later investigations. Let $\Omega_i = \{0,1\}$ and for some finite N consider $\Omega = \Omega_1 \times \cdots \times \Omega_N$. For each i, define probability functions $p_i(1) = q_i$ and $p_i(0) = 1 - q_i$, $q_i \in [0,1]$, and for $\omega = (\omega_1, \dots, \omega_N) \in \Omega$, let $p(\omega) = \prod_i p_i(\omega_i)$. We may interpret this as follows: we toss N coins, where coin i has probability q_i of being heads. The outcome of each toss is independent of all the other tosses.

Exercise$^{\text{(hr)}}$ 1.1.34 (The Birthday Problem)**.** *Assume each day of the year is equally likely to be someone's birthday, and no one is ever born on February 29^{th}. How many people must there be in a room before there is at least a 50% chance that two share a birthday? How many other people must there be before at least one of them shares* your *birthday? Note the two questions have very different answers, because in the first we do not specify beforehand* which *is the shared day, while in the second we do. How many people must be in the room before at least two share a birthday? See also Exercise A.4.8.* Note: in the hint to this problem we show how to approximate the number of people needed before there is a 50% chance that two share a birthday.

Exercise 1.1.35. *Redo the previous problem assuming that there are one-fourth as many people born on February 29^{th} as on any other day.*

Exercise$^{\text{(hr)}}$ 1.1.36. *Two players roll die with k sides, with each side equally likely of being rolled. Player one rolls m dice and player two rolls n dice. If player one's* *highest roll exceeds the highest roll of player two then player one wins, otherwise player two wins. Prove*

$$\text{Prob(Player one wins)} \ = \ \frac{1}{k^{m+n}} \sum_{a=2}^{k} \left[a^m - (a-1)^m\right] \cdot (a-1)^n, \qquad (1.31)$$

which by the integral version of partial summation equals

$$\frac{1}{k^{m+n}} \left[k^m \cdot (k-1)^n - \int_1^k [u]^m \cdot n(u-1)^{n-1} du \right]. \qquad (1.32)$$

If m, n and k are large and of approximately the same size, show

$$\text{Prob(Player one wins)} \ = \ \frac{m}{m+n} - \frac{m}{2(m+n-1)} \frac{n}{k}; \qquad (1.33)$$

note if $m = n = k$ the probability is much less than 50%. See [Mil7] for more details.

1.1.5 Expectation

Definition 1.1.37 (Expected Value)**.** *Consider an outcome space Ω with outcomes ω_i occurring with probabilities $p(\omega_i)$ and a random variable X. The expected value (or mean or average value) of the random variable X is defined by*

$$\overline{X} \ = \ \sum_i X(\omega_i) p(\omega_i). \qquad (1.34)$$

We often write $\mathbb{E}[X]$, read as **the expected value** or **expectation of** X, for $\overline{X}$.

Exercise 1.1.38. *Show the mean of one roll of a fair dice is 3.5. Consider N tosses of a fair coin. Let $X(\omega)$ equal the number of heads in $\omega = (\omega_1, \dots, \omega_N)$. Determine $\mathbb{E}[X]$.*

Remark 1.1.39. Remember we may regard random variables as events; thus it makes sense to talk about the mean value of such events, as the events are real numbers. If we considered an event not arising through a random variable, things would not be as clear. For example, consider $\Omega = \{HH, HT, TH, TT\}$, each with probability $\frac{1}{4}$. We cannot add a head and a tail; however, if we assign a 1 to a head and a 0 to the tail, we need only add numbers.

Exercise 1.1.40. *Consider all finite fair tosses of a coin where all but the last toss are tails (and the last toss is a head). We denote the outcome space by*

$$\Omega \ = \ \{H, TH, TTH, TTTH, \dots\}. \qquad (1.35)$$

Let X be the random variable equal to the number of the toss which is the head. For example, $X(TTH) = 3$. Calculate the probability that the first head is the i^{th} toss. Calculate $\mathbb{E}[X]$.

Definition 1.1.41 (k^{th} Moment)**.** *The k^{th} moment of X is the expected value of x^k. If X is a random variable on an outcome space Ω with events ω_i, we write*

$$\mathbb{E}[X^k] \ = \ \sum_{\omega_i \in \Omega} X(\omega_i)^k \cdot p(\omega_i). \qquad (1.36)$$

Note the first moment is the expected value of X, and the zeroth moment is always 1.

Definition 1.1.42 (Moments of Probability Distributions)**.** *Let $\Omega \subset \mathbb{R}$; thus all events are real numbers, which we shall denote by $x \in \Omega$. Let p be a probability distribution on Ω so that the probability of x is just $p(x)$. We can consider a random variable X with $X(x) = x$; thus the probability that the random variable takes on the value x is $p(x)$. Equivalently we can consider p as a random variable (see Example 1.1.6). We define the k^{th} moment of p by*

$$p_k \;=\; \mathbb{E}[X^k] \;=\; \sum_{x\in\Omega} x^k p(x). \tag{1.37}$$

Similar to how Taylor series coefficients can often determine a "nice" function, a sequence of moments often uniquely determines a probability distribution. We will use such a moment analysis in our Random Matrix Theory investigations in Part **??**; see §**??** for more details.

Exercise 1.1.43. *Prove the zeroth moment of any probability distribution is 1.*

Lemma 1.1.44 (Additivity of the Means)**.** *If X and Y are two random variables on Ω with a probability function p, they induce a joint probability function P with*

$$P(x_i, y_j) \;:=\; \text{Prob}(X(\omega) = x_i, Y(\omega) = y_j). \tag{1.38}$$

Consider the random variable Z, $Z = X + Y$. Then $\mathbb{E}[Z] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proof. First note

$$\text{Prob}(X(\omega) = x_i) \;=\; \sum_j \text{Prob}(X(\omega) = x_i, Y(\omega) = y_j) \;=\; \sum_j P(x_i, y_j). \tag{1.39}$$

Thus the expected value of the random variable X is

$$\mathbb{E}[X] \;=\; \sum_i x_i \sum_j P(x_i, y_j), \tag{1.40}$$

and similarly for the random variable Y. Therefore

$$\begin{aligned}\mathbb{E}[X+Y] &= \sum_{(i,j)} (x_i + y_j) P(x_i, y_j)\\ &= \sum_i \sum_j x_i P(x_i, y_j) + \sum_i \sum_j y_j P(x_i, y_j)\\ &= \sum_i x_i \sum_j P(x_i, y_j) + \sum_j y_j \sum_i P(x_i, y_j)\\ &= \mathbb{E}[X] + \mathbb{E}[Y]. \end{aligned} \tag{1.41}$$

□

The astute reader may notice that some care is needed to interchange the order of summations. If $\sum_i \sum_j |x_i + y_j| p(x_i, y_j) < \infty$, then Fubini's Theorem (Theorem A.2.8) is applicable and we may interchange the summations at will. For an example where the summations cannot be interchanged, see Exercise **??**.

Lemma 1.1.45 (Expectation Is Linear)**.** *Let X_1 through X_N be a finite collection of random variables. Let a_1 through a_N be real constants. Then*

$$\mathbb{E}[a_1X_1 + \cdots + a_NX_N] = a_1\mathbb{E}[X_1] + \cdots + a_N\mathbb{E}[X_N]. \tag{1.42}$$

See §**??** for an application of the linearity of expected values to investigating digits of continued fractions.

Exercise 1.1.46. *Prove Lemma 1.1.45.*

Lemma 1.1.47. *Let X and Y be independent random variables. Then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. From Definition 1.1.32, for all i and j the events $A_i = \{\omega : X(\omega) = x_i\}$ and $B_j = \{\omega : Y(\omega) = y_j\}$ are independent. This implies

$$\text{Prob}(A_i \cap B_j) \ = \ \text{Prob}(A_i)\text{Prob}(B_j) \ = \ p(x_i)q(y_j). \tag{1.43}$$

If $r(x_i, y_j)$ is the probability that the random variable X is x_i and the random variable Y is y_j, then independence implies $r(x_i, y_j) = p(x_i)q(y_j)$ for two probability functions p and q. Thus,

$$\begin{aligned}\mathbb{E}[XY] \ &= \ \sum_i \sum_j x_i y_j r(x_i, y_j) \\ &= \ \sum_i \sum_j x_i y_j p(x_i) q(y_j) \\ &= \ \sum_i x_i p(x_i) \cdot \sum_j y_j q(y_j) \\ &= \ \mathbb{E}[X] \cdot \mathbb{E}[Y]. \end{aligned} \tag{1.44}$$

□

Exercise 1.1.48. *Find two random variables such that $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.*

Exercise 1.1.49 (Two Envelope Problem)**.** *Consider two sealed envelopes; one has X dollars inside and the other has $2X$ dollars, $X > 0$. You are randomly given an envelope — you have an equal likelihood of receiving either. You calculate that you have a 50% chance of having the smaller (larger) amount. Let Y be the amount in your envelope. If you keep this envelope you expect to receive say Y dollars; if you switch your expected value is $.5 \cdot 2Y + .5 \cdot \frac{Y}{2}$, or $1.25Y$. But this is true without ever looking inside the envelope, so you should switch again! What is wrong with the above analysis?*

Exercise$^{\textbf{(hr)}}$ 1.1.50. *Consider a group of m people. We choose a person at random (thus each person is equally likely to be chosen); we do this n times (at each step, each person is equally likely to be chosen). If $n < m$ then clearly there is at least one person whom we haven't chosen. How large must n be so that we have a 50% chance of having chosen everyone at least once? What is the average value of n such that everyone is chosen at least once? See the remarks for applications.*

1.1.6 Variances

The **variance** σ_X^2 and its square root, the **standard deviation** σ_X measure how spread out the values taken on by a random variable are: the larger the variance, the more spread out the distribution.

Definition 1.1.51 (Variance)**.** *Given an outcome space* Ω *with outcomes* ω_i *with probabilities* $p(\omega_i)$ *and a random variable* $X : \Omega \to \mathbb{R}$*, the variance* σ_X^2 *is*

$$\sigma_X^2 \ = \ \sum_i \left(X(\omega_i) - \mathbb{E}[X]\right)^2 p(\omega_i) \ = \ \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]. \tag{1.45}$$

Exercise 1.1.52. *Let* $\Omega_1 = \{0, 25, 50, 75, 100\}$ *with probabilities* $\{.2, .2, .2, .2, .2\}$*, and let* X *be the random variable* $X(\omega) = \omega$*,* $\omega \in \Omega_1$*. Thus* $X(0) = 0, X(25) = 25$*, and so on. Let* Ω_2 *be the same outcome space but with probabilities* $\{.1, .25, .3, .25, .1\}$*, and define* $Y(\omega) = \omega$*,* $\omega \in \Omega_2$*. Calculate the means and the variances of* X *and* Y*.*

For computing variances, instead of (1.45) one often uses

Lemma 1.1.53. *For a random variable* X *we have* $\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$*.*

Proof. Recall $\overline{X} = \mathbb{E}[X]$. Then

$$\begin{aligned} \sigma_X^2 &= \sum_i \left(X_i(\omega) - \mathbb{E}[X]\right)^2 p(\omega_i) \\ &= \sum_i (X_i(\omega)^2 - 2X_i(\omega)\mathbb{E}[X] + \mathbb{E}[X]^2)p(\omega_i) \\ &= \sum_i X_i(\omega)^2 p(\omega_i) - 2\mathbb{E}[X]\sum_i X_i(\omega)p(\omega_i) + \mathbb{E}[X]^2 \sum_i p(\omega_i) \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \end{aligned} \tag{1.46}$$

□

The main result on variances is

Lemma 1.1.54 (Variance of a Sum)**.** *Let* X *and* Y *be two independent random variables on an outcome space* Ω*. Then* $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$*.*

Proof. We use the fact that the expected value of a sum is the sum of expected values (Lemma 1.1.45).

$$\begin{aligned} \sigma_{X+Y}^2 \ &= \ \mathbb{E}[(X+Y)^2] - \mathbb{E}[(X+Y)]^2 \\ &= \ \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \ \left(\mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]\right) - \left(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\right) \\ &= \ \left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \ \sigma_X^2 + \sigma_Y^2 + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right). \end{aligned} \tag{1.47}$$

By Lemma 1.1.47, as X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, completing the proof. □

Let Ω be an outcome space with outcomes ω and a random variable X. For $i \leq N$ let $\Omega_i = \Omega$ and let X_i be the same random variable as X except X_i lives on Ω_i. For example, we could have N rolls with X_i the outcome of the i^{th} roll. We have seen in Lemma 1.1.45 that the mean of the random variable $X_1 + \cdots + X_N$ is $N\mathbb{E}[X]$. What is the variance?

Lemma 1.1.55. *Notation as above,*

$$\sigma_{X_1+\cdots+X_N} \;=\; \sqrt{N}\sigma_X. \tag{1.48}$$

Exercise 1.1.56. *Prove Lemma 1.1.55.*

Lemma 1.1.57. *Given an outcome space Ω with outcomes ω with probabilities $p(\omega)$ and a random variable X. Consider the new random variable $aX + b$. Then*

$$\sigma^2_{aX+b} \;=\; a^2\sigma^2_X. \tag{1.49}$$

Exercise 1.1.58. *Prove 1.1.57.*

Note that if the random variable X has units of meters then the variance σ_X^2 has units of meters^2, and the standard deviation σ_X and the mean $\overline{X}$ have units meters. Thus it is the standard deviation that gives a good measure of the deviations of X around its mean.

There are, of course, alternate measures one can use. For example, one could consider

$$\sum_i (x_i - \overline{X})p(x_i). \tag{1.50}$$

Unfortunately this is a signed quantity, and large positive deviations can cancel with large negatives. In fact, more is true.

Exercise 1.1.59. *Show $\sum_i (x_i - \overline{X})p(x_i) = 0$.*

This leads us to consider

$$\sum_i |x_i - \overline{X}|p(x_i). \tag{1.51}$$

While this has the advantage of avoiding cancellation of errors (as well as having the same units as the events), the absolute value function is not a good function analytically. For example, it is not differentiable. This is primarily why we consider the standard deviation (the square root of the variance).

Exercise 1.1.60 (Method of Least Squares)**.** *Consider the following set of data: for $i \in \{1, \ldots, n\}$, given t_i one observes y_i. Believing that t and y are linearly related, find the best fit straight line. Namely, determine constants a and b that minimize the error (calculated via the variance)*

$$\sum_{i=1}^{n} (y_i - (at_i + b))^2 \;=\; \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2 . \tag{1.52}$$

Hint: *Use multi-variable calculus to find linear equations for a and b, and then solve with linear algebra. If one requires that $a = 0$, show that the b leading to least error is $b = \overline{y} = \frac{1}{n}\sum_i y_i$.*

The method of proof generalizes to the case when one expects y is a linear *combination of N fixed functions. The functions need not be linear; all that is required is that we have a linear combination, say $a_1 f_1(t) + \cdots + a_N f_N(t)$. One then determines the $a_1, \dots, a_N$ that minimize the variance (the sum of squares of the errors) by calculus and linear algebra. If instead of measuring the total error by the squares of the individual error we used another measure (for example, using the absolute value), closed form expressions for the a_i become significantly harder, even in the simple case of fitting a line.*

Exercise 1.1.61. *Consider the best fit line from the Method of Least Squares (Exercise 1.1.60). Is the point $(\overline{x}, \overline{y})$, where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \sum_{i=1}^{n} y_i$, on the best fit line? In other words, does the best fit line go through the "average" point?*

Exercise 1.1.62 (Chebyshev's Inequality)**.** *Let X be a random variable with mean μ and finite variance σ^2. Prove Chebyshev's inequality:*

$$\mathrm{Prob}(|X - \mu| \geq k\sigma) \ \leq \ \frac{1}{k^2}, \tag{1.53}$$

where $\mathrm{Prob}(|X - \mu| \geq a)$ is the probability that X takes on values at least a units from the mean. Chebyshev's theorem holds for all nice distributions, and provides bounds for being far away from the mean (where far is relative to the natural spacing, namely σ).

Exercise 1.1.63. *Use Chebyshev's Theorem to bound the probability of tossing a fair coin 10000 times and observing at least 6000 heads.*

Exercise 1.1.64. *Does there exist a probability distribution such that Chebyshev's Inequality is an equality for all positive integral k?*

If the probability distribution decays sufficiently rapidly we can use the Central Limit Theorem (Theorem 1.4.1) and obtain better estimates than those from Chebyshev's Theorem. See Exercise 1.4.3.

1.2 STANDARD DISTRIBUTIONS

We describe several common probability distributions. Consider the important case when the outcome space $\Omega \subset \mathbb{R}$ and is countable; thus the outcomes are real numbers. Let p be a probability function on Ω. For notational convenience we sometimes extend Ω to all of $\mathbb{R}$ and define the probabilities of the new outcomes as 0.

To each $x \in \mathbb{R}$ we have attached a non-negative number $p(x)$, which is zero except for at most countably many X. We let x_i denote a typical outcome where $p(x) \neq 0$. Similar to calculating the means, variances and higher moments of a random variable, we can compute these quantities for a probability distribution; see Definition 1.1.42. For example, for a discrete probability distribution p the mean is $\sum_i x_i p(x_i)$.

1.2.1 Bernoulli Distribution

Recall the binomial coefficient $\binom{N}{r} = \frac{N!}{r!(N-r)!}$ is the number of ways to choose r objects from N objects when order does not matter; see §A.1.3 for a review of binomial coefficients. Consider n independent repetitions of a process with only two possible outcomes. We typically call one outcome **success** and the other **failure**, the event a **Bernoulli trial**, and a collection of independent Bernoulli trials a **Bernoulli process**. In each Bernoulli trial let there be probability p of success and $q = 1 - p$ of failure. Often we represent a success with 1 and a failure with 0. In §1.2.4 we describe a Bernoulli trial to experimentally determine π!

Exercise 1.2.1. *Consider a Bernoulli trial with random variable X equal to 1 for a success and 0 for a failure. Show $\overline{X} = p$, $\sigma_X^2 = pq$, and $\sigma_X = \sqrt{pq}$. Note X is also an indicator random variable (see Exercise 1.1.14).*

Let Y_N be the number of successes in N trials. Clearly the possible values of Y_N are $\{0, 1, \dots, N\}$. We analyze $p_N(k) = \text{Prob}(Y_N(\omega) = k)$. Here the sample space Ω is all possible sequences of N trials, and the random variable $Y_N : \Omega \to \mathbb{R}$ is given by $Y_N(\omega)$ equals the number of successes in ω.

If $k \in \{0, 1, \dots, N\}$, we need k successes and $N - k$ failures. We do not care what order we have them (i.e., if $k = 4$ and $N = 6$ then $SSFSSF$ and $FSSSSF$ both contribute equally). Each such string of k successes and $N - k$ failures has probability of $p^k \cdot (1-p)^{N-k}$. There are $\binom{N}{k}$ such strings, which implies $p_N(k) = \binom{N}{k} p^k \cdot (1-p)^{N-k}$ if $k \in \{0, 1, \dots, N\}$ and 0 otherwise.

By clever algebraic manipulations, one can directly evaluate the mean $\overline{Y_N}$ and the variance $\sigma_{Y_N}^2$; however, Lemmas 1.1.45 and 1.1.55 allow one to calculate both quantities immediately, once one knows the mean and variance for a single occurrence (see Exercise 1.2.1).

Lemma 1.2.2. *For a Bernoulli process with N trials, each having probability p of success, the expected number of successes is $\overline{Y_N} = Np$ and the variance is $\sigma_{Y_N}^2 = Npq$.*

Lemma 1.2.2 states the expected number of successes is of size Np, and the fluctuations about Np are of size $\sigma_{Y_N}^2 = \sqrt{Npq}$. Thus, if $p = \frac{1}{2}$ and $N = 10^6$, we expect 500,000 successes, with fluctuations on the order of 500. Note how much smaller the fluctuations about the mean are than the mean itself (the mean is of size N, the fluctuations of size $\sqrt{N}$). This is an example of a general phenomenon, which we describe in greater detail in §1.4.

Exercise 1.2.3. *Prove Lemma 1.2.2. Prove the variance is largest when $p = q = \frac{1}{2}$.*

Consider the following problem: Let $\Omega = \{S, FS, FFS, \dots\}$ and let Z be the number of trials before the first success. What is $\overline{Z}$ and σ_Z^2?

First we determine the **Bernoulli distribution** $p(k) = \text{Prob}(Z(\omega) = k)$, the probability that the first success occurs after k trials. Clearly this probability is non-zero only for k a positive integer, in which case the string of results must be

$k-1$ failures followed by 1 success. Therefore

$$p(k) \ = \ \begin{cases} (1-p)^{k-1}\cdot p & \text{if } k\in\{1,2,\dots\} \\ 0 & \text{otherwise.} \end{cases} \tag{1.54}$$

To determine the mean $\overline{Z}$ we must evaluate

$$\overline{Z} \ = \ \sum_{k=1}^{\infty} k(1-p)^{k-1}p \ = \ p\sum_{k=1}^{\infty} kq^{k-1}, \quad 0<q=1-p<1. \tag{1.55}$$

Consider the geometric series

$$f(q) \ = \ \sum_{k=0}^{\infty} q^k \ = \ \frac{1}{1-q}. \tag{1.56}$$

A careful analysis shows we can differentiate term by term if $-1\le q<1$. Then

$$f'(q) \ = \ \sum_{k=0}^{\infty} kq^{k-1} \ = \ \frac{1}{(1-q)^2}. \tag{1.57}$$

Recalling $q=1-p$ and substituting yields

$$\overline{Z} \ = \ p\sum_{k=1}^{\infty} kq^{k-1} \ = \ \frac{p}{\left(1-(1-p)\right)^2} \ = \ \frac{1}{p}. \tag{1.58}$$

Remark 1.2.4. Differentiating under the summation sign is a powerful tool in Probability Theory, and is a common technique for proving such identities. See [Mil4] for more on differentiating identities, where the expected number of alternations between heads and tails in n tosses of a coin with probability p of heads is derived, along with other combinatorial and probability results.

Exercise 1.2.5. *Calculate σ_Z^2.* Hint: *Differentiate $f(q)$ twice.*

Exercise 1.2.6. *Consider the normal distribution with mean 0 and variance σ^2; its density is $f(x;\sigma)=(2\pi\sigma^2)^{-\frac{1}{2}}e^{-x^2/2\sigma^2}$. As $f(x;\sigma)$ integrates to 1, we have*

$$\sigma \ = \ \int_{-\infty}^{\infty} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}}\, dx. \tag{1.59}$$

By differentiating with respect to σ, show the second moment (and hence the variance since the mean is zero) is σ^2. This argument may be generalized (it may be easier to consider the operator $\sigma^3 d/d\sigma$) and yields all even moments of the Gaussian; the $2m^{th}$ moment is $(2m-1)(2m-3)\cdots 3\cdot 1\cdot\sigma^{2m}$ and is often denoted $(2m-1)!!$ (here the double factorial means every other term; thus $7!!=7\cdot 5\cdot 3\cdot 1$ and $6!!=6\cdot 4\cdot 2$).

Exercise 1.2.7. *The even moments of the Gaussian (see Exercise 1.2.6) have an interesting combinatorial meaning. Show that the number of ways of pairing $2m$ objects into m pairs of two elements is $(2m-1)!!$. We shall see these moments again in §**??**, where we study the eigenvalues of real symmetric Toeplitz matrices.*

1.2.2 Poisson Distribution

Divide the unit interval into N equal pieces. Consider N independent Bernoulli trials, one in each subinterval. If the probability of a success is $\frac{\lambda}{N}$, then by Lemma 1.2.2 the expected number of successes is $N \cdot \frac{\lambda}{N} = \lambda$. We consider the limit as $N \to \infty$. We still expect λ successes in each unit interval, but what is the probability of 3λ successes? How long do we expect to wait between successes?

We call this a **Poisson process with parameter** λ. For example, look at the midpoints of the N intervals. At each midpoint we have a Bernoulli trial with probability of success $\frac{\lambda}{N}$ and failure $1 - \frac{\lambda}{N}$. We determine the $N \to \infty$ limits. For fixed N, the probability of *exactly* k successes in a unit interval is

$$\begin{aligned} p_N(k) &= \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N!}{k!(N-k)!} \frac{\lambda^k}{N^k} \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N \cdot (N-1) \cdots (N-k+1)}{N \cdot N \cdots N} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{\lambda}{N}\right)^{-k} \\ &= 1 \cdot \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{k-1}{N}\right) \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N \left(1 - \frac{\lambda}{N}\right)^{-k}. \quad (1.60) \end{aligned}$$

For fixed, finite k and λ, as $N \to \infty$ the first k factors in $p_N(k)$ tend to 1, $\left(1 - \frac{\lambda}{N}\right)^N \to e^{-\lambda}$, and $\left(1 - \frac{\lambda}{N}\right)^{-k} \to 1$ (see §**??** for a review of properties of e). Thus $p_N(k) \to \frac{\lambda^k}{k!} e^{-\lambda}$. We shall see similar calculations as these when we investigate the properties of $x_n = n^k \alpha \bmod 1$ in Chapter **??**.

Using our investigations of Bernoulli trials as a motivation, we are led to the **Poisson Distribution**: Given a parameter λ (interpreted as the expected number of occurrences per unit interval), the probability of k occurrences in a unit interval is $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ for $k \in \{0, 1, 2, \dots\}$. This is a discrete, integer valued process.

Exercise 1.2.8. *Check that $p(k)$ given above is a probability distribution. Namely, show $\sum_{k \geq 0} p(k) = 1$.*

Exercise$^{\text{(h)}}$ 1.2.9. *Calculate the mean and variance for the Poisson Distribution.*

1.2.3 Continuous Distributions

Up to now we have only considered discrete probability distributions. We now study a continuous example. We consider a generalization of a Bernoulli process with λ successes in a unit interval. We divide the real line into subintervals of size $\frac{1}{N}$ and consider a Bernoulli trial at the midpoint of each subinterval with probability $\frac{\lambda}{N}$ of success. Start counting at 0, and let the first success be at X. How is X distributed as $N \to \infty$ (i.e., how long do we expect to wait before seeing the first success)? Denote this distribution by $p_S(x)$.

We have approximately $\frac{x-0}{1/N} = Nx$ midpoints from 0 to X (with N midpoints per unit interval). Let $\lceil y \rceil$ be the smallest integer greater than or equal to y. Then we have $\lceil Nx \rceil$ midpoints, where the results of the Bernoulli trials of the first $\lceil Nx \rceil - 1$ midpoints are all failures and the last is a success. Thus the probability of the first success occurring in an interval of length $\frac{1}{N}$ containing X (with N divisions per unit interval) is

$$p_{N,S}(x) \ = \ \left(1 - \frac{\lambda}{N}\right)^{\lceil Nx \rceil - 1} \cdot \left(\frac{\lambda}{N}\right)^1. \tag{1.61}$$

For N large the above is approximately $e^{-\lambda x}\frac{\lambda}{N}$.

Exercise 1.2.10. *For large N, calculate the size of $N\left(p_{N,s}(x) - e^{-\lambda x}\frac{\lambda}{N}\right)$. Show this difference tends to zero as N tends to infinity.*

Definition 1.2.11 (Continuous Probability Distribution)**.** *We say $p(x)$ is a continuous probability distribution on $\mathbb{R}$ if*

1. *$p(x) \geq 0$ for all $x \in \mathbb{R}$.*
2. *$\int_{\mathbb{R}} p(x)dx = 1$.*
3. *$\text{Prob}(a \leq x \leq b) = \int_a^b p(x)dx$.*

We call $p(x)$ the probability density function or the density; $p(x)dx$ is interpreted as the probability of the interval $[x, x + dx]$.

In the previous example, as $N \to \infty$ we obtain the continuous probability density function

$$p_S(x) \ = \ \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0; \end{cases} \tag{1.62}$$

note $\frac{1}{N}$ is like dx for N large. In the special case of $\lambda = 1$, we get the standard exponential decay, e^{-x}. We will see this distribution in Chapter **??** when we investigate the fractional parts of $n^k\alpha$ (k, α fixed, n varying).

For instance, let $\pi(M)$ be the number of primes that are at most M. The Prime Number Theorem states $\pi(M) = \frac{M}{\log M}$ plus lower order terms. Thus the average spacing between primes around M is about $\log M$. We can model the distribution of primes as a Poisson Process, with parameter $\lambda = \lambda_M = \frac{1}{\log M}$ (this is called the Cramér model). While possible locations of primes (obviously) is discrete (it must be an integer, and in fact the location of primes are not independent), a Poisson model often gives very good heuristics; see for example [Sch].

We often renormalize so that $\lambda = 1$. This is denoted **unit mean spacing**. For example, one can show the M^{th} prime p_M is about $M \log M$, and spacings between primes around p_M is about $\log M$. Then the normalized primes $q_M \approx \frac{p_M}{\log M}$ will have unit mean spacing and $\lambda = 1$.

Example 1.2.12 (**Uniform Distribution on** $[a, b]$)**.** *Let $\Omega = \{x \in \mathbb{R} : a \leq x \leq b\}$. The uniform distribution has probability density function $p(x) = \frac{1}{b-a}$. Note for any $[c, d] \subset [a, b]$,*

$$\text{Prob}\left([c, d]\right) \ = \ \int_c^d p(x)dx \ = \ \frac{d - c}{b - a}. \tag{1.63}$$

The uniform distribution is one of the most common (and best understood!) continuous distributions; the probability of $x \in [c,d] \subset [a,b]$ depends only on the length of the subinterval $[c,d]$.

Example 1.2.13 (Gaussian Distribution)**.** *For $x \in \mathbb{R}$, consider the probability density function $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$. This is called the Gaussian (or normal or bell curve) distribution. By Exercise 1.2.14 it has mean μ and variance σ^2. If $\mu = 0$ and $\sigma^2 = 1$, it is called the standard normal or the standard Gaussian. See §1.4 for more details.*

We sketch the main idea in the proof that the above is a probability distribution. As it is clearly non-negative, we need only show it integrates to one. Consider

$$I \ = \ \int_{-\infty}^{\infty} e^{-x^2} dx. \tag{1.64}$$

Square I, and change from rectangular to polar coordinates, where $dxdy$ becomes $rdrd\theta$:

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2} dx \cdot \int_{-\infty}^{\infty} e^{-y^2} dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2} dxdy \\ &= \int_{0}^{2\pi} d\theta \int_{0}^{\infty} e^{-r^2} rdr \\ &= 2\pi \cdot \left[-\frac{1}{2}e^{-r^2}\right]_0^{\infty} \ = \ \pi. \end{aligned} \tag{1.65}$$

The reason the above works is that while $e^{-x^2}dx$ is hard to integrate, $re^{-r^2}dr$ is easy. Thus $I = \sqrt{\pi}$.

Exercise 1.2.14. *Let $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}$. Prove $\int_{-\infty}^{\infty} p(x)dx = 1$, $\int_{-\infty}^{\infty} xp(x)dx = \mu$ and $\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx = \sigma^2$. This justifies our claim that the Gaussian is a probability distribution with mean μ and variance σ^2.*

Example 1.2.15 (Cauchy Distribution)**.** *Consider*

$$p(x) \ = \ \frac{1}{\pi}\frac{1}{1+x^2}. \tag{1.66}$$

This is a continuous distribution and is symmetric about zero. While we would like to say it therefore has mean zero, the problem is the integral $\int_{-\infty}^{\infty} xp(x)dx$ is not well defined as it depends on how we take the limit. For example,

$$\lim_{A\to\infty}\int_{-A}^{A} xp(x)dx \ = \ 0, \quad \lim_{A\to\infty}\int_{-A}^{2A} xp(x)dx \ = \ \infty. \tag{1.67}$$

Regardless, $p(x)$ has infinite variance. We shall see the Cauchy distribution again in Chapter **??***; see also Exercises* **??** *and* **??**.

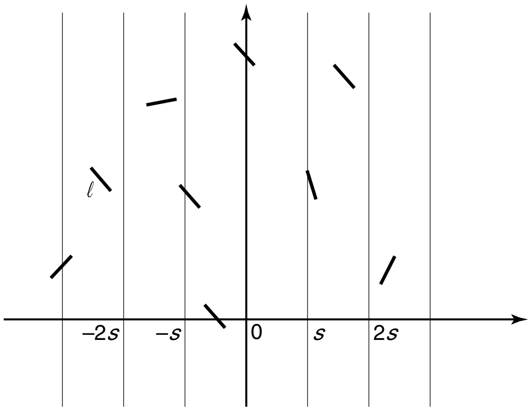

Figure 1.1 Buffon's needle

Exercise 1.2.16. *Prove the Cauchy distribution is a probability distribution by showing*

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx \ = \ 1. \tag{1.68}$$

Show the variance is infinite. See also Exercise **??**.

The Cauchy distribution shows that not all probability distributions have finite moments. When the moments do exist, however, they are a powerful tool for understanding the distribution. The moments play a similar role as coefficients in Taylor series expansions. We use moment arguments to investigate the properties of eigenvalues in Chapters **??** and **??**; see in particular §**??**.

1.2.4 Buffon's Needle and π

We give a nice example of a continuous probability distribution in two dimensions. Consider a collection of infinitely long parallel lines in the plane, where the spacing between any two adjacent lines is s. Let the lines be located at $x = 0, \pm s, \pm 2s, \dots$. Consider a rod of length ℓ where for convenience we assume $\ell < s$. If we were to randomly throw the rod on the plane, what is the probability it hits a line? See Figure 1.1. This question was first asked by Buffon in 1733. For a truly elegant solution which does not use calculus, see [AZ]; we present the proof below as it highlights many of the techniques for investigating probability problems in several variables.

Because of the vertical symmetry we may assume the center of the rod lies on the line $x = 0$, as shifting the rod (without rotating it) up or down will not alter the number of intersections. By the horizontal symmetry, we may assume $-\frac{s}{2} \le x < \frac{s}{2}$. We posit that all values of x are equally likely. As x is continuously distributed, we may add in $x = \frac{s}{2}$ without changing the probability. The probability density function of x is $\frac{dx}{s}$.

Let θ be the angle the rod makes with the x-axis. As each angle is equally likely, the probability density function of θ is $\frac{d\theta}{2\pi}$. We assume that x and θ are chosen independently. Thus the probability density for (x, θ) is $\frac{dx d\theta}{s \cdot 2\pi}$.

The projection of the rod (making an angle of θ with the x-axis) along the x-axis is $\ell \cdot |\cos\theta|$. If $|x| \leq \ell \cdot |\cos\theta|$, then the rod hits exactly one vertical line exactly once; if $x > \ell \cdot |\cos\theta|$, the rod does not hit a vertical line. Note that if $\ell > s$, a rod could hit multiple lines, making the arguments more involved. Thus the probability a rod hits a line is

$$p = \int_{\theta=0}^{2\pi} \int_{x=-\ell\cdot|\cos\theta|}^{\ell\cdot|\cos\theta|} \frac{dx d\theta}{s \cdot 2\pi} = 2\int_{\theta=0}^{2\pi} \frac{\ell \cdot |\cos\theta|}{s} \frac{d\theta}{2\pi} = \frac{2\ell}{\pi s}. \quad (1.69)$$

Exercise 1.2.17. *Show*

$$\frac{1}{2\pi} \int_0^{2\pi} |\cos\theta| d\theta = \frac{2}{\pi}. \quad (1.70)$$

Let A be the random variable which is the number of intersections of a rod of length ℓ thrown against parallel vertical lines separated by $s > \ell$ units. Then

$$A = \begin{cases} 1 & \text{with probability } \frac{2\ell}{\pi s} \\ 0 & \text{with probability } 1 - \frac{2\ell}{\pi s}. \end{cases} \quad (1.71)$$

If we were to throw N rods independently, since the expected value of a sum is the sum of the expected values (Lemma 1.1.45), we expect to observe $N \cdot \frac{2\ell}{\pi s}$ intersections.

Turning this around, let us throw N rods, and let I be the number of observed intersections of the rods with the vertical lines. Then

$$I \approx N \cdot \frac{2\ell}{\pi s} \quad \text{which implies} \quad \pi \approx \frac{N}{I} \cdot \frac{2\ell}{s}. \quad (1.72)$$

The above is an *experimental* formula for π!

Exercise 1.2.18. *Assume we are able to throw the rod randomly as described above, and the N throws are independent. We then have a Bernoulli process with N trials. We have calculated the expected number of successes; using the methods of §1.2.1, calculate the variance (and hence the size of the fluctuations in I). For each N, give the range of values we expect to observe for π.*

1.3 RANDOM SAMPLING

We introduce the notion of **random sampling**. Consider a countable set $\Omega \subset \mathbb{R}$ and a probability function p on Ω; we can extend p to all of $\mathbb{R}$ by setting $p(r) = 0$ if $r \notin \Omega$. Using the probability function p, we can choose elements from $\mathbb{R}$ **at random**. Explicitly, the probability that we choose $\omega \in \Omega$ is $p(\omega)$.

For example, let $\Omega = \{1, 2, 3, 4, 5, 6\}$ with each event having probability $\frac{1}{6}$ (the rolls of a fair die). If we were to roll a fair die N times (for N large), we observe a particular sequence of outcomes. It is natural to assume the rolls are independent of each other. Let X_i denote the outcome of the i^{th} roll. The X_i's all have the same distribution (arising from p). We call the X_i **i.i.d.r.v.** (independent identically distributed random variables), and we say the X_i are a **sample** from the probability distribution p. We say we **randomly sample (with respect to** p**)** $\mathbb{R}$. Often we simply say we have **randomly chosen** N **numbers**.

A common problem is to sample some mathematical or physical process and use the observations to make inferences about the underlying system. For example, we may be given a coin without being told what its probabilities for heads and tails are. We can attempt to infer the probability p of a head by tossing the coin many times, and recoding the outcomes. Let X_i be the outcome of the i^{th} toss (1 for head, 0 for tail). After N tosses we expect to see about Np heads; however, we observe some number, say S_N. Given that we observe S_N heads after N tosses, what is our best guess for p? By Lemma 1.1.45, we guess $p = \frac{S_N}{N}$. It is extremely unlikely that our guess is exactly right. This leads us to a related question: given that we observe S_N heads, can we give a small interval about our best guess where we are extremely confident the true value p lies? The solution is given by the Central Limit Theorem (see §1.4).

Exercise 1.3.1. *For the above example, if* p *is irrational show the best guess can never be correct.*

One can generalize the above to include the important case where p is a continuous distribution. For example, say we wish to investigate the digits of numbers in $[0, 1]$. It is natural to put the uniform distribution on this interval, and choose numbers at random relative to this distribution; we say we choose N numbers randomly with respect to the uniform distribution on $[0, 1]$, or simply we choose N numbers uniformly from $[0, 1]$. Two natural problems are to consider the n^{th} digit in the base 10 expansion and the n^{th} digit in the continued fraction expansion. By observing many choices, we hope to infer knowledge about how these digits are distributed. The first problem is theoretically straightforward. It is not hard to calculate the probability that the n^{th} digit is d; it is just $\frac{1}{10}$. The probabilities of the digits of continued fractions are significantly harder (unlike decimal expansions, *any* positive integer can occur as a digit); see Chapter **??** for the answer.

Exercise 1.3.2 (Important for Computational Investigations)**.** *For any continuous distribution* p *on* $\mathbb{R}$*, the probability we chose a number in* $[a, b]$ *is* $\int_a^b p(x)dx$*. If we were to choose* N *numbers,* N *large, then we expect approximately* $N \int_a^b p(x)dx$ *to be in* $[a, b]$*. Often computers have built in random number generators for certain continuous distributions, such as the standard Gaussian or the uniform, but not for less common ones. Show if one can randomly choose numbers from the uniform distribution, one can use this to randomly choose from any distribution.* Hint: *Use* $C_p(x) = \int_{-\infty}^{x} p(x)dx$*, the* ***Cumulative Distribution Function*** *of* p *(see also §**??**); it is the probability of observing a number at most* x*.*

Remark 1.3.3. The observant reader may notice a problem with sampling from a continuous distribution: the probability of choosing any particular real number is zero, but some number is chosen! One explanation is that, fundamentally, we cannot choose numbers from a continuous probability distribution. For example, if we use computers to choose our numbers, all computers can do is a finite number of manipulations of 0's and 1's; thus, they can only choose numbers from a countable (actually finite) set. The other interpretation of the probability of any $r \in \mathbb{R}$ is zero is that, while at each stage some number is chosen, no number is ever chosen twice. Thus, in some sense, any number we explicitly write down is "special." See also Exercise 1.1.49, where the resolution is that one cannot choose numbers uniformly on all of $(0, \infty)$.

For our investigations, we approximate continuous distributions by discrete distributions with many outcomes. From a practical point of view, this suffices for many experiments; however, one should note that while theoretically we can write statements such as "choose a real number uniformly from $[0, 1]$," we can never actually do this.

1.4 THE CENTRAL LIMIT THEOREM

We close our introduction to probability with a statement of *the* main theorem about the behavior of a sum of independent events. We give a proof in an important special case in §1.4.2 and sketch the proof in general in §**??**. For more details and weaker conditions, see [Bi, CaBe, Fe]. We discuss applications of the Central Limit Theorem to determining whether or not numerical experiments support a conjecture in Chapter 2.

1.4.1 Statement of the Central Limit Theorem

Let X_i ($i \in \{1, \dots, N\}$) be independent identically distributed random variables (i.i.d.r.v.) as in §1.3, all sampled from the same probability distribution p with mean μ and variance σ^2; thus $\mathbb{E}[X_i] = \mu$ and $\sigma^2_{X_i} = \sigma^2$ for all i. Let $S_N = \sum_{i=1}^N X_i$. We are interested in the distribution of the random variable S_N as $N \to \infty$. As each X_i has expected value μ, by Lemma 1.1.45 $\mathbb{E}[S_N] = N\mu$. We now consider a more refined question: how is S_N distributed about $N\mu$? The Central Limit Theorem answers this, and tells us what the correct scale is to study the fluctuations about $N\mu$.

Theorem 1.4.1 (Central Limit Theorem)**.** *For $i \in \{1, \dots, N\}$, let X_i be i.i.d.r.v. with mean μ, finite variance σ^2 and finite third moment. Let $S_N = X_1 + \cdots + X_N$. As $N \to \infty$*

$$\text{Prob}(S_N \in [\alpha, \beta]) \ \sim \ \frac{1}{\sqrt{2\pi\sigma^2 N}} \int_\alpha^\beta e^{-(t-\mu N)^2/2\sigma^2 N} dt. \tag{1.73}$$

In other words, the distribution of S_N converges to a Gaussian with mean μN and variance $\sigma^2 N$. We may re-write this as

$$\lim_{N\to\infty} \text{Prob}\left(\frac{S_N - \mu N}{\sqrt{\sigma^2 N}} \in [a, b]\right) \ = \ \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt. \tag{1.74}$$

Here $Z_N = \frac{S_N - \mu N}{\sqrt{\sigma^2 N}}$ converges to a Gaussian with mean 0 and variance 1.

The probability density $\frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}$ is the **standard Gaussian**. It is *the* universal curve of probability. Note how robust the Central Limit Theorem is: it does not depend on fine properties of the X_i, just that they all have the same distributions and finite variance (and a bit more). While this is true in most situations, it fails in some cases such as sampling from a Cauchy distribution (see Exercise **??** for another limit theorem which can handle such cases). Sometimes it is important to know how rapidly Z_N is converging to the Gaussian. The rate of convergence *does* depend on the higher moments; see §**??** and [Fe].

Exercise 1.4.2. *The Central Limit Theorem gives us the correct scale to study fluctuations. For example, say we toss a fair coin N times (hence $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{4}$). We expect S_N to be about $\frac{N}{2}$. Find values of a and b such that the probability of $S_N - N\mu \in [a\sqrt{N}/2, b\sqrt{N}/2]$ converges to 95% (resp., 99%). For large N, show for any fixed $\delta > 0$ that the probability of $S_N - N\mu \in [aN^{\frac{1}{2}+\delta}/2, bN^{\frac{1}{2}+\delta}/2]$ tends to zero. Thus we expect to observe half of the tosses as heads, and we expect deviations from one-half to be of size $2/\sqrt{N}$.*

Exercise 1.4.3. *Redo Exercise 1.1.63 using the Central Limit Theorem and compare the two bounds.*

Exercise 1.4.4. *For $S_N = X_1 + \cdots + X_N$, calculate the variance of $Z_N = \frac{S_N - \mu N}{\sqrt{\sigma^2 N}}$; this shows $\sqrt{\sigma^2 N}$ is the correct scale to investigate fluctuations of S_N about μN.*

One common application of the Central Limit Theorem is to test whether or not we are sampling the X_i independently from a fixed probability distribution with mean μ and known standard deviation σ (if the standard deviation is not known, there are other tests which depend on methods to estimate σ). Choose N numbers randomly from what we expect has mean μ. We form S_N as before and investigate $\frac{S_N - \mu N}{\sqrt{\sigma^2 N}}$. As $S_N = \sum_{i=1}^{N} X_i$, we expect S_N to be of size N. If the X_i are not drawn from a distribution with mean μ, then $S_N - N\mu$ will also be of size N. Thus, $\frac{S_N - N\mu}{\sqrt{\sigma^2 N}}$ will be of size $\sqrt{N}$ if the X_i are not drawn from something with mean μ. If, however, the X_i are from sampling a distribution with mean μ, the Central Limit Theorem states that $\frac{S_N - N\mu}{\sqrt{\sigma^2 N}}$ will be of size 1. See Chapter 2 for more details and Exercise **??** for an alternate sampling statistic.

Finally, we note that the Central Limit Theorem is an example of the **Philosophy of Square Root Cancellation**: the sum is of size N, but the deviations are of size $\sqrt{N}$. We have already seen examples of such cancellation in Remark **??** and §**??**, and will see more in our investigations of writing integers as the sum of primes (see §**??**).

1.4.2 Proof for Bernoulli Processes

We sketch the proof of the Central Limit Theorem for Bernoulli Processes where the probability of success is $p = \frac{1}{2}$. Consider the random variable X that is 1 with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$ (for example, tosses of a fair coin; the advantage of making a tail -1 is that the mean is zero). Note the mean of X is

$\overline{X} = 0$, the variance is $\sigma_X^2 = 1$ (as we have $1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2}$) and the standard deviation is $\sigma_X = 1$.

Let $X_1, \dots, X_{2N}$ be independent identically distributed random variables, distributed as X (it simplifies the expressions to consider an even number of tosses). Consider $S_{2N} = X_1 + \cdots + X_{2N}$. Its mean is zero and its variance is $2N$, and we expect fluctuations of size $\sqrt{2N}$. We show that for N large the distribution of S_{2N} is approximately normal. We need

Lemma 1.4.5 (Stirling's Formula)**.** *For n large,*

$$n! \;=\; n^n e^{-n}\sqrt{2\pi n}\,(1 + O(1/n))\,. \tag{1.75}$$

For a proof, see [WW]. We show (1.75) is a reasonable approximation. It is often easier to analyze a product by converting it to a sum; this is readily accomplished by taking logarithms. We have

$$\log n! \;=\; \sum_{k=1}^{n} \log k \;\approx\; \int_1^n \log t dt \;=\; (t\log t - t)|_1^n. \tag{1.76}$$

Thus $\log n! \approx n\log n - n$, or $n! \approx n^n e^{-n}$.

We now consider the distribution of S_{2N}. We first note that the probability that $S_{2N} = 2k+1$ is zero. This is because S_{2N} equals the number of heads minus the number of tails, which is always even: if we have k heads and $2N-k$ tails then S_{2N} equals $2N - 2k$.

The probability that S_{2N} equals $2k$ is just $\binom{2N}{N+k}(\frac{1}{2})^{N+k}(\frac{1}{2})^{N-k}$. This is because for S_{2N} to equal $2k$, we need $2k$ more 1's (heads) than -1's (tails), and the number of 1's and -1's add to $2N$. Thus we have $N+k$ heads (1's) and $N-k$ tails (-1's). There are 2^{2N} strings of 1's and -1's, $\binom{2N}{N+k}$ have exactly $N+k$ heads and $N-k$ tails, and the probability of each string is $(\frac{1}{2})^{2N}$. We have written $(\frac{1}{2})^{N+k}(\frac{1}{2})^{N-k}$ to show how to handle the more general case when there is a probability p of heads and $1-p$ of tails.

We use Stirling's Formula to approximate $\binom{2N}{N+k}$. After elementary algebra we find

$$\begin{aligned}\binom{2N}{N+k} \;&\approx\; \frac{(2N)^{2N}}{(N+k)^{N+k}(N-k)^{N-k}}\sqrt{\frac{N}{\pi(N+k)(N-k)}} \\ &=\; \frac{2^{2N}}{\sqrt{\pi N}}\frac{1}{(1+\frac{k}{N})^{N+\frac{1}{2}+k}(1-\frac{k}{N})^{N+\frac{1}{2}-k}}.\end{aligned} \tag{1.77}$$

We would like to use $\left(1+\frac{w}{N}\right)^N \approx e^w$ from §**??**; unfortunately, we must be a little more careful as the values of k we consider grow with N. For example, we might believe that $(1+\frac{k}{N})^N \to e^k$ and $(1-\frac{k}{N})^N \to e^{-k}$, so these factors cancel. As k is small relative to N we may ignore the factors of $\frac{1}{2}$, and then say

$$\left(1+\frac{k}{N}\right)^k \;=\; \left(1+\frac{k}{N}\right)^{N\cdot\frac{k}{N}} \;\to\; e^{k^2/N}; \tag{1.78}$$

similarly, $(1-\frac{k}{N})^{-k} \to e^{k^2/N}$. Thus we would claim (*and we shall see later in Lemma 1.4.6 that this claim is in error!*) that

$$\left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \;\to\; e^{2k^2/N}. \tag{1.79}$$

We show that $\left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\ \left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \to e^{k^2/N}$. The importance of this calculation is that it highlights how crucial rates of convergence are. While it is true that the main terms of $(1 \pm \frac{k}{N})^N$ are $e^{\pm k}$, the error terms (in the convergence) are quite important, and yield large secondary terms when k is a power of N. What happens here is that the secondary terms from these two factors reinforce each other. Instead of using $\left(1+\frac{w}{N}\right)^N \approx e^w$ from §**??**, it is better to take the logarithms of the two factors, Taylor expand, and then exponentiate. This allows us to better keep track of the error terms.

An immediate consequence of Chebyshev's inequality (see Exercise 1.1.62) is that we need only study k where $|k|$ is at most $N^{\frac{1}{2}+\epsilon}$. This is because the standard deviation of S_{2N} is $\sqrt{2N}$. Specifically, see Exercise 1.4.8 for a proof that given any $\epsilon > 0$, the probability of observing a k with $|k| \gg N^{\frac{1}{2}+\epsilon}$ is negligible. Thus it suffices to analyze the probability that $S_{2N} = 2k$ for $|k| \le N^{\frac{1}{2}+\frac{1}{9}}$.

Lemma 1.4.6. *For any $\epsilon \le \frac{1}{9}$, for $N \to \infty$ with $k \ll N^{\frac{1}{2}+\epsilon}$, we have*

$$\left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\ \left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \ \to\ e^{k^2/N}e^{O(N^{-1/6})}. \tag{1.80}$$

Proof. Recall that for $|x| < 1$,

$$\log(1+x) \ =\ \sum_{n=1}^{\infty} \frac{(-1)^{n+1}x^n}{n}. \tag{1.81}$$

As we are assuming $k \ll N^{\frac{1}{2}+\epsilon}$, note that any term below of size k^2/N^2, k^3/N^2 or k^4/N^3 will be negligible. Thus we have

$$\begin{aligned}
P_{k,N} &= \left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\ \left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \\
\log P_{k,N} &= \left(N+\frac{1}{2}+k\right)\log\left(1+\frac{k}{N}\right) + \left(N+\frac{1}{2}-k\right)\log\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \\
&= \left(N+\frac{1}{2}+k\right)\left(\frac{k}{N}-\frac{k^2}{2N^2}+O\left(\frac{k^3}{N^3}\right)\right) \\
&\quad + \left(N+\frac{1}{2}-k\right)\left(-\frac{k}{N}-\frac{k^2}{2N^2}+O\left(\frac{k^3}{N^3}\right)\right) \\
&= \frac{2k^2}{N} - 2\left(N+\frac{1}{2}\right)\frac{k^2}{2N^2} + O\left(\frac{k^3}{N^2}+\frac{k^4}{N^3}\right) \\
&= \frac{k^2}{N} + O\left(\frac{k^2}{N^2}+\frac{k^3}{N^2}+\frac{k^4}{N^3}\right).
\end{aligned} \tag{1.82}$$

As $k \ll N^{\frac{1}{2}+\epsilon}$, for $\epsilon < \frac{1}{9}$ the big-Oh term is dominated by $N^{-1/6}$, and we finally obtain that

$$P_{k,N} \ =\ e^{k^2/N}e^{O\left(N^{-1/6}\right)}, \tag{1.83}$$

which completes the proof. □

Combining Lemma 1.4.6 with (1.77) yields

$$\binom{2N}{N+k}\frac{1}{2^{2N}} \approx \frac{1}{\sqrt{\pi N}}\, e^{-k^2/N}. \tag{1.84}$$

The proof of the central limit theorem in this case is completed by some simple algebra. We are studying $S_{2N} = 2k$, so we should replace k^2 with $(2k)^2/4$. Similarly, since the variance of S_{2N} is $2N$, we should replace N with $(2N)/2$. We find

$$\text{Prob}(S_{2N} = 2k) \;=\; \binom{2N}{N+k}\frac{1}{2^{2N}} \approx \frac{2}{\sqrt{2\pi \cdot (2N)}}\, e^{-(2k)^2/2(2N)}. \tag{1.85}$$

Remember S_{2N} is never odd. The factor of 2 in the numerator of the normalization constant above reflects this fact, namely the contribution from the probability that S_{2N} is even is twice as large as we would expect, because it has to account for the fact that the probability that S_{2N} is odd is zero. Thus the above looks like a Gaussian with mean 0 and variance $2N$. For N large such a Gaussian is slowly varying, and integrating from $2k$ to $2k+2$ is basically $2/\sqrt{2\pi(2N)} \cdot \exp -(2k)^2/2(2N)$.

Exercise 1.4.7. *Use the integral test to bound the error in* (1.76)*, and then use that to bound the error in the estimate of* $n!$.

Exercise 1.4.8. *Prove the standard deviation of* S_{2N} *is* $\sqrt{2N}$*. Use this and Chebyshev's inequality (Exercise 1.1.62) to prove*

$$\text{Prob}(|S_{2N}| \geq N^{\epsilon} \cdot \sqrt{2N}) \;\leq\; \frac{1}{N^{2\epsilon}}, \tag{1.86}$$

which implies that it suffices to study values of k *with* $k \ll N^{\frac{1}{2}+\epsilon}$.

Exercise 1.4.9. *Prove* (1.81).

Exercise 1.4.10. *Can you generalize the above arguments to handle the case when* $p \neq \frac{1}{2}$.

Chapter Two

Applications of Probability: Benford's Law and Hypothesis Testing

The Gauss-Kuzmin Theorem (Theorem **??**) tells us that the probability that the millionth digit of a randomly chosen continued fraction expansion is k is approximately $q_k = \log_2\left(1 + \frac{1}{k(k+2)}\right)$. What if we choose N algebraic numbers, say the cube roots of N consecutive primes: how often do we expect to observe the millionth digit equal to k? If we believe that algebraic numbers other than rationals and quadratic irrationals satisfy the Gauss-Kuzmin Theorem, we expect to observe $q_k N$ digits equal to k, and probably fluctuations on the order of $\sqrt{N}$. If we observe M digits equal to k, how confident are we (as a function of M and N, of course) that the digits are distributed according to the Gauss-Kuzmin Theorem? This leads us to the subject of **hypothesis testing**: if we assume some process has probability p of success, and we observe M successes in N trials, does this provide support for or against the hypothesis that the probability of success is p?

We develop some of the theory of hypothesis testing by studying a concrete problem, the distribution of the first digit of certain sequences. In many problems (for example, 2^n base 10), the distribution of the first digit is given by Benford's Law, described below. We first investigate situations where we can easily prove the sequences are Benford, and then discuss how to analyze data in harder cases where the proofs are not as clear (such as the famous $3x + 1$ problem). The error analysis is, of course, the same as the one we would use to investigate whether or not the digits of the continued fraction expansions of algebraic numbers satisfy the Gauss-Kuzmin Theorem. In the process of investigating Benford's Law, we encounter equidistributed sequences (Chapter **??**), logarithmic probabilities (similar to the Gauss-Kuzmin probabilities in Chapter **??**), and Poisson Summation (Chapter **??**), as well as many of the common problems in statistical testing (such as non-independent events and multiple comparisons).

2.1 BENFORD'S LAW

While looking through tables of logarithms in the late 1800s, Newcomb noticed a surprising fact: certain pages were significantly more worn out than others. People were looking up numbers whose logarithm started with 1 more frequently than other digits. In 1938 Benford [Ben] observed the same digit bias in a variety of phenomenon. See [Hi1, Rai] for a description and history, [Hi2, BBH, KonMi, LaSo, MN] for recent results, [Knu] for connections between Benford's law and rounding errors in computer calculations and [Nig1, Nig2] for applications of Benford's Law by the IRS to detect corporate tax fraud!

A sequence of positive numbers $\{x_n\}$ is **Benford (base** b**)** if the probability of observing the first digit of x_n in base b is j is $\log_b\left(1+\frac{1}{j}\right)$. More precisely,

$$\lim_{N\to\infty} \frac{\#\{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\}}{N} \ = \ \log_b\left(1+\frac{1}{j}\right). \tag{2.1}$$

Note that $j \in \{1,\dots,b-1\}$. This is a probability distribution as one of the $b-1$ events must occur, and the total probability is

$$\sum_{j=1}^{b-1} \log_b\left(1+\frac{1}{j}\right) \ = \ \log_b \prod_{j=1}^{b-1}\left(1+\frac{1}{j}\right) \ = \ \log_b \prod_{j=1}^{b-1} \frac{j+1}{j} \ = \ \log_b b \ = \ 1. \tag{2.2}$$

It is possible to be Benford to some bases but not others; we show the first digit of 2^n is Benford base 10, but clearly it is not Benford base 2 as the first digit is always 1. For many processes, we obtain a sequence of points, and the distribution of the first digits are Benford. For example, consider the **3x+1 problem**. Let a_0 be any positive integer, and consider the sequence where

$$a_{n+1} \ = \ \begin{cases} 3a_n+1 & \text{if } a_n \text{ is odd} \\ a_n/2 & \text{if } a_n \text{ is even.} \end{cases} \tag{2.3}$$

For example, if $a_0 = 13$, we have

$$\begin{aligned} 13 &\longrightarrow 40 \longrightarrow 20 \longrightarrow 10 \longrightarrow 5 \longrightarrow 16 \longrightarrow 8 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \\ &\longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \cdots. \end{aligned} \tag{2.4}$$

An alternate definition is to remove as many powers of two as possible in one step. Thus

$$a_{n+1} \ = \ \frac{3a_n+1}{2^k}, \tag{2.5}$$

where k is the largest power of 2 dividing $3a_n+1$. It is conjectured that for *any* a_0, eventually the sequence becomes $4 \to 2 \to 1 \to 4 \cdots$ (or in the alternate definition $1 \to 1 \to 1 \cdots$). While this is known for all $a_0 \le 2^{60}$, the problem has resisted numerous attempts at proofs (Kakutani has described the problem as a conspiracy to slow down mathematical research because of all the time spent on it). See [Lag1, Lag2] for excellent surveys of the problem. How do the first digits behave for a_0 large? Do numerical simulations support the claim that this process is Benford? Does it matter which definition we use?

Exercise 2.1.1. *Show the Benford probabilities* $\log_{10}\left(1+\frac{1}{j}\right)$ *for* $j \in \{1,\dots,9\}$ *are irrational. What if instead of base ten we work in base* d *for some integer* d*?*

Exercise 2.1.2. *Below we use the definition of the* $3x+1$ *map from* (2.5). *Show there are arbitrarily large integers* N *such that if* $a_0 = N$ *then* $a_1 = 1$. *Thus, infinitely often, one iteration is enough to enter the repeating cycle. More generally, for each positive integer* k *does there exist arbitrarily large integers* N *such that if* $a_0 = N$ *then* $a_j > 1$ *for* $j < k$ *and* $a_k = 1$*?*

2.2 BENFORD'S LAW AND EQUIDISTRIBUTED SEQUENCES

As we can write any positive x as b^u for some u, the following lemma shows that it suffices to investigate $u \bmod 1$:

Lemma 2.2.1. *The first digits of b^u and b^v are the same in base b if and only if $u \equiv v$ mod 1.*

Proof. We prove one direction as the other is similar. If $u \equiv v \bmod 1$, we may write $v = u + m$, $m \in \mathbb{Z}$. If

$$b^u \;=\; u_k b^k + u_{k-1}b^{k-1} + \cdots + u_0 + u_{-1}b^{-1} + \cdots, \tag{2.6}$$

then

$$\begin{aligned} b^v \;&=\; b^{u+m} \\ &=\; b^u \cdot b^m \\ &=\; (u_k b^k + u_{k-1}b^{k-1} + \cdots + u_0 + u_{-1}b^{-1} + \cdots)b^m \\ &=\; u_k b^{k+m} + \cdots + u_0 b^m + u_{-1}b^{m-1} + \cdots . \end{aligned} \tag{2.7}$$

Thus the first digits of each are u_k, proving the claim. □

Exercise 2.2.2. *Prove the other direction of the if and only if.*

Consider the unit interval $[0, 1)$. For $j \in \{1, \dots, b\}$, define p_j by

$$b^{p_j} = j \quad \text{or equivalently} \quad p_j \;=\; \log_b j. \tag{2.8}$$

For $j \in \{1, \dots, b-1\}$, let

$$I_j^{(b)} \;=\; [p_j, p_{j+1}) \;\subset [0, 1). \tag{2.9}$$

Lemma 2.2.3. *The first digit of b^y base b is j if and only if y mod $1 \in I_j^{(b)}$.*

Proof. By Lemma 2.2.1 we may assume $y \in [0, 1)$. Then $y \;\in I_j^{(b)} = [p_j, p_{j+1})$ if and only if $b^{p_j} \;\le\; y \;<\; b^{p_{j+1}}$, which from the definition of p_j is equivalent to $j \le b^y < j + 1$, proving the claim. □

The following theorem shows that the exponentials of equidistributed sequences (see Definition **??**) are Benford.

Theorem 2.2.4. *If $y_n = \log_b x_n$ is equidistributed mod 1 then x_n is Benford (base b).*

Proof. By Lemma 2.2.3,

$$\begin{aligned} &\{n \le N : y_n \bmod 1 \in [\log_b j, \log_b(j+1))\} \\ &\qquad =\; \{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\}\,. \end{aligned} \tag{2.10}$$

Therefore

$$\begin{aligned} &\lim_{N\to\infty} \frac{\#\left\{n \le N : y_n \bmod 1 \in [\log_b j, \log_b(j+1))\right\}}{N} \\ &\qquad =\; \lim_{N\to\infty} \frac{\#\left\{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\right\}}{N}. \end{aligned} \tag{2.11}$$

If y_n is equidistributed, then the left side of (2.11) is $\log_b\left(1 + \frac{1}{j}\right)$ which implies x_n is Benford base b. □

Remark 2.2.5. One can extend the definition of Benford's Law from statements concerning the distribution of the first digit to the distribution of the first k digits. With such an extension, Theorem 2.2.4 becomes $y_n = \log_b x_n \bmod 1$ is equidistributed if and only if x_n is Benford base b. See [KonMi] for details.

Let $\{x\} = x - [x]$ denote the fractional part of x, where $[x]$ as always is the greatest integer at most x. In Theorem **??** we prove that for $\alpha \notin \mathbb{Q}$ the fractional parts of $n\alpha$ are equidistributed modulo 1. From this and Theorem 2.2.4, it immediately follows that geometric series are Benford (modulo the irrationality condition):

Theorem 2.2.6. *Let $x_n = ar^n$ with $\log_b r \notin \mathbb{Q}$. Then x_n is Benford (base b).*

Proof. Let $y_n = \log_b x_n = n \log_b r + \log_b a$. As $\log_b r \notin \mathbb{Q}$, by Theorem **??** the fractional parts of y_n are equidistributed. Exponentiating by b, we obtain that x_n is Benford (base b) by Theorem 2.2.4. □

Theorem 2.2.6 implies that 2^n is Benford base 10, but not surprisingly that it is not Benford base 2.

Exercise 2.2.7. *Do the first digits of e^n follow Benford's Law? What about $e^n + e^{-n}$?*

2.3 RECURRENCE RELATIONS AND BENFORD'S LAW

We show many sequences defined by recurrence relations are Benford. For more on recurrence relations, see Exercise **??**. The interested reader should see [BrDu, NS] for more on the subject.

2.3.1 Recurrence Preliminaries

We consider recurrence relations of length k:

$$a_{n+k} \ = \ c_1 a_{n+k-1} + \cdots + c_k a_n, \tag{2.12}$$

where $c_1, \dots, c_k$ are fixed real numbers. If the characteristic polynomial

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - \cdots - c_{k-1} r - c_k \ = \ 0 \tag{2.13}$$

has k distinct roots $\lambda_1, \dots, \lambda_k$, there exist k numbers $u_1, \dots, u_k$ such that

$$a_n \ = \ u_1 \lambda_1^n + \cdots + u_k \lambda_k^n, \tag{2.14}$$

where we have ordered the roots so that $|\lambda_1| \ge \cdots \ge |\lambda_k|$.

For the Fibonacci numbers $k = 2$, $c_1 = c_2 = 1$, $u_1 = -u_2 = \frac{1}{\sqrt{5}}$, and $\lambda_1 = \frac{1+\sqrt{5}}{2}$, $\lambda_2 = \frac{1-\sqrt{5}}{2}$ (see Exercise **??**). If $|\lambda_1| = 1$, we do not expect the first digit of a_n to be Benford (base b). For example, if we consider

$$a_n \ = \ 2a_{n-1} - a_{n-2} \tag{2.15}$$

with initial values $a_0 = a_1 = 1$, every $a_n = 1$! If we instead take $a_0 = 0$, $a_1 = 1$, we get $a_n = n$. See [Kos] for many interesting occurrences of Fibonacci numbers and recurrence relations.

2.3.2 Recurrence Relations Are Benford

Theorem 2.3.1. *Let a_n satisfy a recurrence relation of length k with k distinct real roots. Assume $|\lambda_1| \neq 1$ with $|\lambda_1|$ the largest absolute value of the roots. Further, assume the initial conditions are such that the coefficient of λ_1 is non-zero. If $\log_b |\lambda_1| \notin \mathbb{Q}$, then a_n is Benford (base b).*

Proof. By assumption, $u_1 \neq 0$. For simplicity we assume $\lambda_1 > 0$, $\lambda_1 > |\lambda_2|$ and $u_1 > 0$. Again let $y_n = \log_b x_n$. By Theorem 2.2.4 it suffices to show y_n is equidistributed mod 1. We have

$$\begin{aligned} x_n &= u_1\lambda_1^n + \cdots + u_n\lambda_k^n \\ x_n &= u_1\lambda_1^n \left[1 + O\left(\frac{ku\lambda_2^n}{\lambda_1^n}\right)\right], \end{aligned} \tag{2.16}$$

where $u = \max_i |u_i| + 1$ (so $ku > 1$ and the big-Oh constant is 1). As $\lambda_1 > |\lambda_2|$, we "borrow" some of the growth from λ_1^n; this is a very useful technique. Choose a small ϵ and an n_0 such that

1. $|\lambda_2| < \lambda_1^{1-\epsilon}$;
2. for all $n > n_0$, $\frac{(ku)^{1/n}}{\lambda_1^\epsilon} < 1$, which then implies $\frac{ku}{\lambda_1^{n\epsilon}} = \left(\frac{(ku)^{1/n}}{\lambda_1^\epsilon}\right)^n$.

As $ku > 1$, $(ku)^{1/n}$ is decreasing to 1 as n tends to infinity. Note $\epsilon > 0$ if $\lambda_1 > 1$ and $\epsilon < 0$ if $\lambda_1 < 1$. Letting

$$\beta = \frac{(ku)^{1/n_0}}{\lambda_1^\epsilon}\frac{|\lambda_2|}{\lambda_1^{1-\epsilon}} < 1, \tag{2.17}$$

we find that the error term above is bounded by β^n for $n > n_0$, which tends to 0. Therefore

$$\begin{aligned} y_n &= \log_b x_n \\ &= \log_b(u_1\lambda_1^n) + O\left(\log_b(1+\beta^n)\right) \\ &= n\log_b \lambda_1 + \log_b u_1 + O(\beta^n), \end{aligned} \tag{2.18}$$

where the big-Oh constant is bounded by C say. As $\log_b \lambda_1 \notin \mathbb{Q}$, the fractional parts of $n\log_b \lambda_1$ are equidistributed modulo 1, and hence so are the shifts obtained by adding the fixed constant $\log_b u_1$.

We need only show that the error term $O(\beta^n)$ is negligible. It is possible for the error term to change the first digit; for example, if we had 999999 (or 1000000), then if the error term contributes 2 (or -2), we would change the first digit base 10. However, for n sufficiently large, the error term will change a vanishingly small number of first digits. Say $n\log_b \lambda_1 + \log_b u_1$ exponentiates base b to first digit j, $j \in \{1, \dots, b-1\}$. This means

$$n\log_b \lambda_1 + \log_b u_1 \in I_j^{(b)} = [p_{j-1}, p_j). \tag{2.19}$$

The error term is at most $C\beta^n$ and y_n exponentiates to a different first digit than $n\log_b \lambda_1 + \log_b u_1$ only if one of the following holds:

1. $n \log_b \lambda_1 + \log_b u_1$ is within $C\beta^n$ of p_j, and adding the error term pushes us to or past p_j;
2. $n \log_b \lambda_1 + \log_b u_1$ is within $C\beta^n$ of p_{j-1}, and adding the error term pushes us before p_{j-1}.

The first set is contained in $[p_j - C\beta^n, p_j)$, of length $C\beta^n$. The second is contained in $[p_{j-1}, p_{j-1} + C\beta^n)$, also of length $C\beta^n$. Thus the length of the interval where $n \log_b \lambda_1 + \log_b u_1$ and y_n could exponentiate base b to different first digits is of size $2C\beta^n$. If we choose N sufficiently large then for all $n > N$ we can make these lengths arbitrarily small. As $n \log_b \lambda_1 + \log_b u_1$ is equidistributed modulo 1, we can control the size of the subsets of $[0, 1)$ where $n \log_b \lambda_1 + \log_b u_1$ and y_n disagree. The Benford behavior (base b) of x_n now follows in the limit. □

Exercise 2.3.2. *Weaken the conditions of Theorem 2.3.1 as much as possible. What if several roots equal λ_1? What does a general solution to* (2.12) *look like now? What if λ_1 is negative? Can anything be said if there are complex roots?*

Exercise(hr) 2.3.3. *Consider the recurrence relation $a_{n+1} = 5a_n - 8a_{n-1} + 4a_{n-2}$. Show there is a choice of initial conditions such that the coefficient of λ_1 (a largest root of the characteristic polynomial) is non-zero but the sequence does not satisfy Benford's Law.*

Exercise(hr) 2.3.4. *Assume all the roots of the characteristic polynomial are distinct, and let λ_1 be the largest root in absolute value. Show for almost all initial conditions that the coefficient of λ_1 is non-zero, which implies that our assumption that $u_1 \neq 0$ is true most of the time.*

2.4 RANDOM WALKS AND BENFORD'S LAW

Consider the following (colorful) problem: A drunk starts off at time zero at a lamppost. Each minute he stumbles with probability p one unit to the right and with probability $q = 1 - p$ one unit to the left. Where do we expect the drunk to be after N tosses? This is known as a **Random Walk**. By the Central Limit Theorem (Theorem 1.4.1), his distribution after N tosses is well approximated by a Gaussian with mean $1 \cdot pN + (-1) \cdot (1 - p)N = (2p - 1)N$ and variance $p(1 - p)N$. For more details on Random Walks, see [Re].

For us, a **Geometric Brownian Motion** is a process such that its logarithm is a Random Walk (see [Hu] for complete statements and applications). We show below that the first digits of Geometric Brownian Motions are Benford. In [KonSi] the $3x + 1$ problem is shown to be an example of Geometric Brownian Motion. For heuristic purposes we use the first definition of the $3x + 1$ map, though the proof is for the alternate definition. We have two operators: T_3 and T_2, with $T_3(x) = 3x + 1$ and $T_2(x) = \frac{x}{2}$. If a_n is odd, $3a_n + 1$ is even, so T_3 must always be followed by T_2. Thus, we have really have two operators T_2 and $T_{3/2}$, with $T_{3/2}(x) = \frac{3x+1}{2}$. If we assume each operator is equally likely, half the time we go from $x \to \frac{3}{2}x + 1$, and half the time to $\frac{1}{2}x$.

If we take logarithms, $\log x$ goes to $\log \frac{3}{2}x = \log x + \log \frac{3}{2}$ half the time and $\log \frac{1}{2}x = \log x + \log \frac{1}{2}$ the other half. Hence on average we send $\log x \to \log x + \frac{1}{2}\log \frac{3}{4}$. As $\log \frac{3}{4} < 0$, on average our sequence is decreasing (which agrees with the conjecture that eventually we reach $4 \to 2 \to 1$). Thus we might expect our sequence to look like $\log x_k = \log x + \frac{k}{2}\log \frac{3}{4}$. As $\log \frac{3}{4} \notin \mathbb{Q}$, its multiples are equidistributed modulo 1, and thus when we exponentiate we expect to see Benford behavior. Note, of course, that this is simply a heuristic, suggesting we might see Benford's Law. A better heuristic is sketched in Exercise 2.4.1.

While we can consider Random Walks or Brownian Motion with non-zero means, for simplicity below we assume the means are zero. Thus, in the example above, $p = \frac{1}{2}$.

Exercise$^{\text{(hr)}}$ 2.4.1. *Give a better heuristic for the Geometric Brownian Motion of the $3x+1$ map by considering the alternate definition: $a_{n+1} = \frac{3a_n+1}{2^k}$, where $2^k||3x+1$. In particular, calculate the expected value of $\log a_{n+1}$. To do so, we need to estimate the probability $k = \ell$ for each $\ell \in \{1,2,3,\dots\}$; note $k \neq 0$ as for x odd, $3x+1$ is always even and thus divisible by at least one power of 2. Show it is reasonable to assume that $\mathrm{Prob}(k=\ell) = 2^{-\ell}$.*

2.4.1 Needed Gaussian Integral

Consider a sequence of Gaussians G_σ with mean 0 and variance σ^2, with $\sigma^2 \to \infty$. The following lemma shows that for any $\delta > 0$ as $\sigma \to \infty$ almost all of the probability is in the interval $[-\sigma^{1+\delta}, \sigma^{1+\delta}]$. We will use this lemma to show that it is enough to investigate Gaussians in the range $[-\sigma^{1+\delta}, \sigma^{1+\delta}]$.

Lemma 2.4.2.

$$\frac{2}{\sqrt{2\pi\sigma^2}}\int_{\sigma^{1+\delta}}^{\infty} e^{-x^2/2\sigma^2}dx \ \ll\ e^{-\sigma^{2\delta}/2}. \tag{2.20}$$

Proof. Change the variable of integration to $w = \frac{x}{\sigma\sqrt{2}}$. Denoting the above integral by I, we find

$$I \ =\ \frac{2}{\sqrt{2\pi\sigma^2}}\int_{\sigma^\delta/\sqrt{2}}^{\infty} e^{-w^2}\cdot\sigma\sqrt{2}dw \ =\ \frac{2}{\sqrt{\pi}}\int_{\sigma^\delta/\sqrt{2}}^{\infty} e^{-w^2}dw. \tag{2.21}$$

The integrand is monotonically decreasing. For $w \in \left[\frac{\sigma^\delta}{\sqrt{2}}, \frac{\sigma^\delta}{\sqrt{2}}+1\right]$, the integrand is bounded by substituting in the left endpoint, and the region of integration is of length 1. Thus,

$$\begin{aligned} I &< 1 \cdot \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}+1}^{\infty} e^{-w^2} dw \\ &= \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-(u+1)^2} du \\ &= \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-u^2} e^{-2u} e^{-1} du \\ &< \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{e\sqrt{\pi}} e^{-\sigma^{2\delta}/2} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-2u} du \\ &< \frac{2(e+1)}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} \\ &< 4e^{-\sigma^{2\delta}/2}. \end{aligned} \tag{2.22}$$

□

Exercise 2.4.3. *Prove a similar result for intervals of the form* $[-\sigma g(\sigma), \sigma g(\sigma)]$ *where* $g(\sigma)$ *is a positive increasing function and* $\lim_{\sigma\to\infty} g(\sigma) = +\infty$*.*

2.4.2 Geometric Brownian Motions Are Benford

We investigate the distribution of digits of processes that are Geometric Brownian Motions. By Theorem 2.2.4 it suffices to show that the Geometric Brownian Motion converges to being equidistributed modulo 1. Explicitly, we have the following: after N iterations, by the Central Limit Theorem the expected value converges to a Gaussian with mean 0 and variance proportional to $\sqrt{N}$. We must show that the Gaussian with growing variance is equidistributed modulo 1.

For convenience we assume the mean is 0 and the variance is $N/2\pi$. This corresponds to a fair coin where for each head (resp., tail) we move $\frac{1}{\sqrt{4\pi}}$ units to the right (resp., left). By the Central Limit Theorem the probability of being x units to the right of the origin after N tosses is asymptotic to

$$p_N(x) \; = \; \frac{e^{-\pi x^2/N}}{\sqrt{N}}. \tag{2.23}$$

For ease of exposition, we assume that rather than being asymptotic to a Gaussian, the distribution is a Gaussian. For our example of flipping a coin, this cannot be true. If every minute we flip a coin and record the outcome, after N minutes there are 2^N possible outcomes, a finite number. To each of these we attach a number equal to the excess of heads to tails. There are technical difficulties in working with discrete probability distributions; thus we study instead continuous processes such that at time N the probability of observing x is given by a Gaussian with mean 0 and variance $N/2\pi$. For complete details see [KonMi].

Theorem 2.4.4. *As* $N \to \infty$, $p_N(x) = \frac{e^{-\pi x^2/N}}{\sqrt{N}}$ *becomes equidistributed modulo* 1.

Proof. For each N we calculate the probability that for $x \in \mathbb{R}$, $x \bmod 1 \in [a,b] \subset [0,1)$. This is

$$\int_{\substack{x=-\infty \\ x \bmod 1 \in [a,b]}}^{\infty} p_N(x)dx \ = \ \frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx. \tag{2.24}$$

We need to show the above converges to $b-a$ as $N \to \infty$. For $x \in [a,b]$, standard calculus (Taylor series expansions, see §A.2.3) gives

$$e^{-\pi(x+n)^2/N} \ = \ e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N}\, e^{-n^2/N}\right). \tag{2.25}$$

We claim that in (2.24) it is sufficient to restrict the summation to $|n| \le N^{5/4}$. The proof is immediate from Lemma 2.4.2: we increase the integration by expanding to $x \in [0,1]$, and then trivially estimate. Thus, up to negligible terms, all the contribution is from $|n| \le N^{5/4}$.

In §**??** we prove the Poisson Summation formula, which in this case yields

$$\frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} e^{-\pi n^2/N} \ = \ \sum_{n\in\mathbb{Z}} e^{-\pi n^2 N}. \tag{2.26}$$

The beauty of Poisson Summation is that it converts one infinite sum with *slow* decay to another sum with *rapid* decay; because of this, Poisson Summation is an extremely useful technique for a variety of problems. The exponential terms on the left of (2.26) are all of size 1 for $n \le \sqrt{N}$, and do not become small until $n \gg \sqrt{N}$ (for instance, once $n > \sqrt{N}\log N$, the exponential terms are small for large N); however, almost all of the contribution on the right comes from $n = 0$. The power of Poisson Summation is it often allows us to approximate well long sums with short sums. We therefore have

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{|n|\le N^{5/4}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx \\
&\quad= \ \frac{1}{\sqrt{N}} \sum_{|n|\le N^{5/4}} \int_{x=a}^{b} \left[e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N} e^{-n^2/N}\right)\right] dx \\
&\quad= \ \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \sum_{n=0}^{N^{5/4}} \frac{n+1}{\sqrt{N}} e^{-\pi(n/\sqrt{N})^2}\right) \\
&\quad= \ \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \int_{w=0}^{N^{3/4}} (w+1) e^{-\pi w^2} \sqrt{N} dw\right) \\
&\quad= \ \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(N^{-1/2}\right).
\end{aligned} \tag{2.27}$$

By Lemma 2.4.2 we can extend all sums to $n \in \mathbb{Z}$ in (2.27) with negligible error. We now apply Poisson Summation and find that up to lower order terms,

$$\frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx \ \approx \ (b-a) \cdot \sum_{n\in\mathbb{Z}} e^{-\pi n^2 N}. \tag{2.28}$$

For $n = 0$ the right hand side of (2.28) is $b-a$. For all other n, we trivially estimate the sum:

$$\sum_{n\neq 0} e^{-\pi n^2 N} \ \le\ 2\sum_{n\geq 1} e^{-\pi n N} \ \le\ \frac{2e^{-\pi N}}{1-e^{-\pi N}}, \tag{2.29}$$

which is less than $4e^{-\pi N}$ for N sufficiently large. □

We can interpret the above arguments as follows: for each N, consider a Gaussian $p_N(x)$ with mean 0 and variance $N/2\pi$. As $N \to \infty$ for each x (which occurs with probability $p_N(x)$) the first digit of 10^x converges to the Benford base 10 probabilities.

Remark 2.4.5. The above framework is very general and applicable to a variety of problems. In [KonMi] it is shown that these arguments can be used to prove Benford behavior in discrete systems such as the $3x+1$ problem as well as continuous systems such as the absolute values of the Riemann zeta function (and any "good" L-function) near the critical line! For these number theory results, the crucial ingredients are Selberg's result (near the critical line, $\log|\zeta(s+it)|$ for $t \in [T, 2T]$ converges to a Gaussian with variance tending to infinity in T) and estimates by Hejhal on the rate of convergence. For the $3x+1$ problem the key ingredients are the structure theorem (see [KonSi]) and the approximation exponent of Definition **??**; see [LaSo] for additional results on Benford behavior of the $3x+1$ problem.

2.5 STATISTICAL INFERENCE

Often we have reason to believe that some process occurs with probability p of success and $q = 1-p$ of failure. For example, consider the $3x+1$ problem. Choose a large a_0 and look at the first digit of the a_n's. There is reason to believe the distribution of the first digits is given by Benford's Law for most a_0 as $a_0 \to \infty$. We describe how to test this and similar hypotheses. We content ourselves with describing one simple test; the interested reader should consult a statistics textbook (for example, [BD, CaBe, LF, MoMc]) for the general theory and additional applications.

2.5.1 Null and Alternative Hypotheses

Suppose we think some population has a parameter with a certain value. If the population is small, it is possible to investigate every element; in general this is not possible.

For example, say the parameter is how often the millionth decimal or continued fraction digit is 1 in two populations: all rational numbers in $[0,1)$ with denominator at most 5, and all real numbers in $[0,1)$. In the first, there are only 10 numbers, and it is easy to check them all. In the second, as there are infinitely many numbers, it is impossible to numerically investigate each. What we do in practice is we sample a large number of elements (say N elements) in $[0,1)$, and calculate the average value of the parameter for this sample.

We thus have two **populations**, the **underlying population** (in the second case, all numbers in $[0, 1)$), and the **sample population** (in this case, the N sampled elements).

Our goal is to test whether or not the underlying population's parameter has a given value, say p. To this end, we want to compare the sample population's value to p. The **null hypothesis**, denoted H_0, is the claim that there is no difference between the sample population's value and the underlying population's value; the **alternative hypothesis**, denoted H_a, is the claim that there is a difference between the sample population's value and the underlying population's value.

When we analyze the data from the sample population, either we reject the null hypothesis, or we fail to reject the null hypothesis. It is important to note that we *never* prove the null or alternative hypothesis is true or false. We are always rejecting or failing to reject the null hypothesis, we are never accepting it. If we flip a coin 100 times and observe all heads, this does not mean the coin is not fair: it is possible the coin is fair but we had a very unusual sample (though, of course, it is extremely unlikely).

We now discuss how to test the null hypothesis. Our main tool is the Central Limit Theorem. This is just one of many possible inference tests; we refer the reader to [BD, CaBe, LF, MoMc] for more details.

2.5.2 Bernoulli Trials and the Central Limit Theorem

Assume we have some process where we expect a probability p of observing a given value. For example, if we choose numbers uniformly in $[0, 1)$ and look at the millionth decimal digit, we believe that the probability this digit is 1 is $\frac{1}{10}$. If we look at the continued fraction expansion, by Theorem **??** the probability that the millionth digit is 1 is approximately $\log_2 \frac{4}{3}$. What if we restrict to algebraic numbers? What is the probability the millionth digit (decimal or continued fraction expansion) equals 1?

In general, once we formalize our conjecture we test it by choosing N elements from the population independently at random (see §1.3). Consider the claim that a process has probability p of success. We have N independent Bernoulli trials (see §1.2.1). The null hypothesis is the claim that p percent of the sample are a success. Let S_N be the number of successes; if the null hypothesis is correct, by the Central Limit Theorem (see §1.4) we expect S_N to have a Gaussian distribution with mean pN and variance pqN (see Exercise 1.2.1 for the calculations of the mean and variance of a Bernoulli process). This means that if we were to look at many samples with N elements, on average each sample would have $pN \pm O(\sqrt{pqN})$ successes. We calculate the probability of observing a difference $|S_N - pN|$ as large or larger than a. This is given by the area under the Gaussian with mean pN and variance pqN:

$$\frac{1}{\sqrt{2\pi pqN}} \int_{|s-pN|\geq a} e^{-(s-pN)^2/2pqN} ds. \tag{2.30}$$

If this integral is small, it is extremely unlikely that we choose N independent trials from a process with probability p of success and we reject the null hypothesis; if the integral is large, we do not reject the null hypothesis, and we have support for our claim that the underlying process does have probability p of success.

Unfortunately, the Gaussian is a difficult function to integrate, and we would need to tabulate these integrals for *every* different pair of mean and variance. It is easier, therefore, to renormalize and look at a new statistic which should also be Gaussian, but with mean 0 and variance 1. The advantage is that we need only tabulate *one* special Gaussian, the standard normal.

Let $Z = \frac{S_N - pN}{\sqrt{pqN}}$. This is known as the **z-statistic**. If S_N's distribution is a Gaussian with mean pN and variance pqN, note Z will be a Gaussian with mean 0 and variance 1.

Exercise 2.5.1. *Prove the above statement about the distribution of* z.

Let

$$I(a) \ = \ \frac{1}{\sqrt{2\pi}} \int_{|z| \geq a} e^{-z^2/2} dz, \tag{2.31}$$

the area under the standard normal (mean 0, standard deviation 1) that is at least a units from the mean. We consider different **confidence intervals**. If we were to randomly choose a number z from such a Gaussian, what is the probability (as a function of a) that z is at most a units from the mean? Approximately 68% of the time $|z| \leq 1$ ($I(1) \approx .32$), approximately 95% of the time $z \leq 1.96$ ($I(1.96) \approx .05$), and approximately 99% of the time $|z| \leq 2.57$ ($I(2.57) = .01$). In other words, there is only about a 1% probability of observing $|z| \geq 2.57$. If $|z| \geq 2.57$, we have strong evidence against the hypothesis that the process occurs with probability p, and we would be reasonably confident in rejecting the null hypothesis; of course, it is possible we were unlucky and obtained an unrepresentative set of data (but it is extremely unlikely that this occurred; in fact, the probability is at most 1%).

Remark 2.5.2. For a Gaussian with mean μ and standard deviation σ, the probability that $|X - \mu| \leq \sigma$ is approximately $.68$. Thus if X is drawn from a normal with mean μ and standard deviation σ, then approximately 68% of the time $\mu \in [x - \sigma, x + \sigma]$ (where x is the observed value of the random variable X).

To test the claim that some process occurs with probability p, we observe N independent trials, calculate the z-statistic, and see how likely it is to observe $|Z|$ that large or larger. We give two examples below.

2.5.3 Digits of the $3x + 1$ Problem

Consider again the $3x + 1$ problem. Choose a large integer a_0, and look at the iterates: $a_1, a_2, a_3, \dots$. We study how often the first digit of terms in the sequence equal $d \in \{1, \dots, 9\}$. We can regard the first digit of a term as a Bernoulli trial with a success (or 1) if the first digit is d and a failure (or 0) otherwise. If the distribution of digits is governed by Benford's Law, the theoretical prediction is that the fraction of the first digits that equal d is $p = \log_{10}(\frac{d+1}{d})$. Assume there are N terms in our sequence (before we hit the pattern $4 \to 2 \to 1 \to 4 \cdots$), and say M of them have first digit d. For what M does this experiment provide support that the digits follow Benford's Law?

Exercise 2.5.3. *The terms in the sequence generated by a_0 are not independent, as a_{n+1} is determined by a_n. Show that if the first digit of a_n is 2 then the first digit of a_{n+1} cannot be a 2.*

The above exercise shows that the first digit of the terms *cannot* be considered independent Bernoulli trials. As the sequence is completely determined by the first term, this is not surprising. If we look at an enormous number of terms, however, these effects "should" average out. Another possible experiment is to look at the first digit of the millionth term for N different a_0's.

Let $a_0 = 333\dots333$ be the integer that is 10,000 threes. There are 177,857 terms in the sequence before we hit $4 \to 2 \to 1$. The following data comparing the number of first digits equal to d to the Benford predictions are from [Min]:

digit	observed	predicted	variance	z-statistic	$I(z)$
1	53425	53540	193.45	-0.596	0.45
2	31256	31310	160.64	-0.393	0.31
3	22257	22220	139.45	0.257	0.21
4	17294	17230	124.76	0.464	0.36
5	14187	14080	113.88	0.914	0.63
6	11957	11900	105.40	0.475	0.36
7	10267	10310	98.57	-0.480	0.37
8	9117	9090	92.91	0.206	0.16
9	8097	8130	88.12	-0.469	0.36

As the values of the z-statistics are all small (well below 1.96 and 2.57), the above table provides evidence that the first digits in the $3x + 1$ problem follow Benford's Law, and we would not reject the null hypothesis for any of the digits. If we had obtained large z-statistics, say 4, we would reject the null hypothesis and doubt that the distribution of digits follow Benford's Law.

Remark 2.5.4 (Important)**.** One must be very careful when analyzing all the digits. Once we know how many digits are in $\{1, \dots, 8\}$, then the number of 9's is forced: these are not nine independent tests, and a different statistical test (a chi-square test with eight degrees of freedom) should be done. Our point here is not to write a treatise on statistical inference, but merely highlight some of the tools and concepts. See [BD, CaBe, LF, MoMc] for more details, and [Mil5] for an amusing analysis of a baseball problem involving chi-square tests.

Additionally, if we have many different experiments, then "unlikely" events should happen. For example, if we have 100 different experiments we would not be surprised to see an outcome which only has a 1% chance of occurring (see Exercise 2.5.5). Thus, if there are many experiments, the confidence intervals need to be adjusted. One common method is the Bonferroni adjustment method for multiple comparisons. See [BD, MoMc].

Exercise 2.5.5. *Assume for each trial there is a 95% chance of observing the fraction of first digits equal to 1 is in $[\log_{10} 2 - 1.96\sigma, \log_{10} 2 + 1.96\sigma]$ (for some*

σ). If we have 10 independent trials, what is the probability that all *the observed percentages are in this interval? If we have 14 independent trials?*

Remark 2.5.6. How does one calculate with $10,000$ digit numbers? Such large numbers are greater than the standard number classes (int, long, double) of many computer programming languages. The solution is to represent numbers as arrays. To go from a_n to $3a_n + 1$, we multiply the array by 3, carrying as needed, and then add 1; we leave space-holding zeros at the start of the array. For example,

$$3 \cdot [0, \dots, 0, 0, 5, 6, 7] \;=\; [0, \dots, 0, 1, 7, 0, 1]. \tag{2.32}$$

We need only do simple operations on the array. For example, $3 \cdot 7 = 21$, so the first entry of the product array is 1 and we carry the 2 for the next multiplication. We must also compute $a_n/2$ if a_n is even. Note this is the same as $5a_n$ divided by 10. The advantage of this approach is that it is easy to calculate $5a_n$, and as a_n is even, the last digit of $5a_n$ is zero, hence array division by 10 is trivial.

Exercise 2.5.7. *Consider the first digits of the $3x + 1$ problem (defined as in* (2.3)*) in base 6. Choose a large integer a_0, and look at the iterates $a_1, a_2, a_3, \dots$. As $a_0 \to \infty$, is the distribution of digits Benford base 6?*

Exercise 2.5.8 (Recommended)**.** *Here is another variant of the $3x + 1$ problem:*

$$a_{n+1} \;=\; \begin{cases} 3a_n + 1 & \text{if } a_n \text{ is odd} \\ a_n/2^k & \text{if } a_n \text{ is even and } 2^k || a_n; \end{cases} \tag{2.33}$$

$2^k || a_n$ means 2^k divides a_n, but 2^{k+1} does not. Consider the distribution of first digits of this sequence for various a_0. What is the null hypothesis? Do the data support the null hypothesis, or the alternative hypothesis? Do you think these numbers also satisfy Benford's Law? What if instead we define

$$a_{n+1} \;=\; \frac{3a_n + 1}{2^k}, \quad 2^k || a_n. \tag{2.34}$$

2.5.4 Digits of Continued Fractions

Let us test the hypothesis that the digits of algebraic numbers are given by the Gauss-Kuzmin Theorem (Theorem **??**). Let us look at how often the 1000^{th} digit equals 1. By the Gauss-Kuzmin Theorem this should be approximately $\log_2 \frac{4}{3}$. Let p_n be the n^{th} prime. In the continued fraction expansions of $\sqrt[3]{p_n}$ for $n \in \{100000, 199999\}$, exactly 41565 have the 1000^{th} digit equal to 1. Assuming we have a Bernoulli process with probability of success (a digit of 1) of $p = \log_2 \frac{4}{3}$, the z-statistic is $.393$. As the z-statistic is small (95% of the time we expect to observe $|z| \le 1.96$), we do not reject the null hypothesis, and we have obtained evidence supporting the claim that the probability that the 1000^{th} digit is 1 is given by the Gauss-Kuzmin Theorem. See Chapter **??** for more detailed experiments on algebraic numbers and the Gauss-Kuzmin Theorem.

2.6 SUMMARY

We have chosen to motivate our presentation of statistical inference by investigating the first digits of the $3x + 1$ problem, but of course the methods apply to a variety of problems. Our main tool is the Central Limit Theorem: if we have a process with probability p (resp., $q = 1 - p$) of success (resp., failure), then in N independent trials we expect about pN successes, with fluctuations of size $\sqrt{pqN}$. To test whether or not the underlying probability is p we formed the z-statistic: $\frac{S_N - pN}{\sqrt{pqN}}$, where S_N is the number of successes observed in the N trials.

If the process really does have probability p of success, then by the Central Limit Theorem the distribution of S_N is approximately a Gaussian with mean pN and standard deviation $\sqrt{pqN}$, and we then expect the z-statistic to be of size 1. If, however, the underlying process occurs not with probability p but p', then we expect S_N to be approximately a Gaussian with mean $p'N$ and standard deviation $\sqrt{p'q'N}$. We now expect the z-statistic to be of size $\frac{(p'-p)N}{\sqrt{p'q'N}}$. This is of size $\sqrt{N}$, much larger than 1.

We see the z-statistic is very sensitive to $p' - p$: if p' is differs from p, for large N we quickly observe large values of z. Note, of course, that statistical tests can only provide compelling evidence in favor or against a hypothesis, never a proof.

Appendix A

Analysis Review

A.1 PROOFS BY INDUCTION

Assume for each positive integer n we have a statement $P(n)$ which we desire to show is true. $P(n)$ is true for all positive integers n if the following two statements hold:

- **Basis Step:** $P(1)$ is true;
- **Inductive Step**: whenever $P(n)$ is true, $P(n+1)$ is true.

This technique is called **Proof by Induction**, and is a very useful method for proving results; we shall see many instances of this in this appendix and Chapter **??** (indeed, throughout much of the book). The reason the method works follows from basic logic. We assume the following two sentences are true:

$$\begin{aligned} P(1) &\text{ is true} \\ \forall n \geq 1, P(n) \text{ is true implies } P(n+1) &\text{ is true.} \end{aligned} \tag{A.1}$$

Set $n = 1$ in the second statement. As $P(1)$ is true, and $P(1)$ implies $P(2)$, $P(2)$ must be true. Now set $n = 2$ in the second statement. As $P(2)$ is true, and $P(2)$ implies $P(3)$, $P(3)$ must be true. And so on, completing the proof. Verifying the first statement the **basis step** and the second the **inductive step**. In verifying the inductive step, note we assume $P(n)$ is true; this is called the **inductive assumption**. Sometimes instead of starting at $n = 1$ we start at $n = 0$, although in general we could start at any n_0 and then prove for all $n \geq n_0$, $P(n)$ is true.

We give three of the more standard examples of proofs by induction, and one false example; the first example is the most typical.

A.1.1 Sums of Integers

Let $P(n)$ be the statement

$$\sum_{k=1}^{n} k \ = \ \frac{n(n+1)}{2}. \tag{A.2}$$

Basis Step: $P(1)$ is true, as both sides equal 1.
Inductive Step: Assuming $P(n)$ is true, we must show $P(n+1)$ is true. By the

inductive assumption, $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$. Thus

$$\begin{aligned} \sum_{k=1}^{n+1} k &= (n+1) + \sum_{k=1}^{n} k \\ &= (n+1) + \frac{n(n+1)}{2} \\ &= \frac{(n+1)(n+1+1)}{2}. \end{aligned} \tag{A.3}$$

Thus, given $P(n)$ is true, then $P(n+1)$ is true.

Exercise A.1.1. *Prove*

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}. \tag{A.4}$$

Find a similar formula for the sum of k^3. See also Exercise **??**.

Exercise A.1.2. *Show the sum of the first n odd numbers is n^2, i.e.,*

$$\sum_{k=1}^{n} (2k-1) = n^2. \tag{A.5}$$

Remark A.1.3. We define the empty sum to be 0, and the empty product to be 1. For example, $\sum_{n \in \mathbb{N}, n<0} 1 = 0$.

See [Mil4] for an alternate derivation of sums of powers that does not use induction.

A.1.2 Divisibility

Let $P(n)$ be the statement 133 divides $11^{n+1} + 12^{2n-1}$.

Basis Step: A straightforward calculation shows $P(1)$ is true: $11^{1+1} + 12^{2-1} = 121 + 12 = 133$.
Inductive Step: Assume $P(n)$ is true, i.e., 133 divides $11^{n+1} + 12^{2n-1}$. We must show $P(n+1)$ is true, or that 133 divides $11^{(n+1)+1} + 12^{2(n+1)-1}$. But

$$\begin{aligned} 11^{(n+1)+1} + 12^{2(n+1)-1} &= 11^{n+1+1} + 12^{2n-1+2} \\ &= 11 \cdot 11^{n+1} + 12^2 \cdot 12^{2n-1} \\ &= 11 \cdot 11^{n+1} + (133+11)12^{2n-1} \\ &= 11\left(11^{n+1} + 12^{2n-1}\right) + 133 \cdot 12^{2n-1}. \end{aligned} \tag{A.6}$$

By the inductive assumption 133 divides $11^{n+1} + 12^{2n-1}$; therefore, 133 divides $11^{(n+1)+1} + 12^{2(n+1)-1}$, completing the proof.

Exercise A.1.4. *Prove 4 divides $1 + 3^{2n+1}$.*

A.1.3 The Binomial Theorem

We prove the Binomial Theorem. First, recall that

Definition A.1.5 (Binomial Coefficients)**.** *Let n and k be integers with $0 \leq k \leq n$. We set*

$$\binom{n}{k} \;=\; \frac{n!}{k!(n-k)!}. \tag{A.7}$$

Note that $0! = 1$ and $\binom{n}{k}$ is the number of ways to choose k objects from n (with order not counting).

Lemma A.1.6. *We have*

$$\binom{n}{k} \;=\; \binom{n}{n-k}, \quad \binom{n}{k} + \binom{n}{k-1} \;=\; \binom{n+1}{k}. \tag{A.8}$$

Exercise A.1.7. *Prove Lemma A.1.6.*

Theorem A.1.8 (The Binomial Theorem)**.** *For all positive integers n we have*

$$(x+y)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k. \tag{A.9}$$

Proof. We proceed by induction.
Basis Step: For $n = 1$ we have

$$\sum_{k=0}^{1} \binom{1}{k} x^{1-k} y^k \;=\; \binom{1}{0} x + \binom{1}{1} y \;=\; (x+y)^1. \tag{A.10}$$

Inductive Step: Suppose

$$(x+y)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k. \tag{A.11}$$

Then using Lemma A.1.6 we find that

$$\begin{aligned}
(x+y)^{n+1} \;&=\; (x+y)(x+y)^n \\
&=\; (x+y) \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \\
&=\; \sum_{k=0}^{n} \binom{n}{k} x^{n+1-k} y^k + \binom{n}{k} x^{n-k} y^{k+1} \\
&=\; x^{n+1} + \sum_{k=1}^{n} \left\{ \binom{n}{k} + \binom{n}{k-1} \right\} x^{n+1-k} y^k + y^{n+1} \\
&=\; \sum_{k=0}^{n+1} \binom{n+1}{k} x^{n+1-k} y^k.
\end{aligned} \tag{A.12}$$

This establishes the induction step, and hence the theorem. □

A.1.4 False Proofs by Induction

Consider the following: let $P(n)$ be the statement that in any group of n people, everyone has the same name. We give a (false!) proof by induction that $P(n)$ is true for all n!

Basis Step: Clearly, in any group with just 1 person, every person in the group has the same name.

Inductive Step: Assume $P(n)$ is true, namely, in any group of n people, everyone has the same name. We now prove $P(n+1)$. Consider a group of $n+1$ people:

$$\{1, 2, 3, \dots, n-1, n, n+1\}. \tag{A.13}$$

The first n people form a group of n people; by the inductive assumption, they all have the same name. So, the name of 1 is the same as the name of 2 is the same as the name of 3 $\dots$ is the same as the name of n.

Similarly, the last n people form a group of n people; by the inductive assumption they all have the same name. So, the name of 2 is the same as the name of 3 $\dots$ is the same as the name of n is the same as the name of $n+1$. Combining yields everyone has the same name! Where is the error?

If $n = 4$, we would have the set $\{1, 2, 3, 4, 5\}$, and the two sets of 4 people would be $\{1, 2, 3, 4\}$ and $\{2, 3, 4, 5\}$. We see that persons 2, 3 and 4 are in both sets, providing the necessary link.

What about smaller n? What if $n = 1$? Then our set would be $\{1, 2\}$, and the two sets of 1 person would be $\{1\}$ and $\{2\}$; there is no overlap! The error was that we assumed n was "large" in our proof of $P(n) \Rightarrow P(n+1)$.

Exercise A.1.9. *Show the above proof that $P(n)$ implies $P(n+1)$* is *correct for $n \geq 2$, but fails for $n = 1$.*

Exercise A.1.10. *Similar to the above, give a false proof that any sum of k integer squares is an integer square, i.e., $x_1^2 + \cdots + x_n^2 = x^2$. In particular, this would prove all positive integers are squares as $m = 1^2 + \cdots + 1^2$.*

Remark A.1.11. There is no such thing as *Proof By Example*. While it is often useful to check a special case and build intuition on how to tackle the general case, checking a few examples is not a proof. For example, because $\frac{16}{64} = \frac{1}{4}$ and $\frac{19}{95} = \frac{1}{5}$, one might think that in dividing two digit numbers if two numbers on a diagonal are the same one just cancels them. If that were true, then $\frac{12}{24}$ should be $\frac{1}{4}$. Of course this is *not* how one divides two digit numbers!

A.2 CALCULUS REVIEW

We briefly review some of the results from Differential and Integral Calculus. We recall some notation: $[a, b] = \{x : a \leq x \leq b\}$ is the set of all x between a and b, including a and b; $(a, b) = \{x : a < x < b\}$ is the set of all x between a and b, not including the endpoints a and b. For a review of continuity see §A.3.

A.2.1 Intermediate Value Theorem

Theorem A.2.1 (Intermediate Value Theorem (IVT)). *Let f be a continuous function on $[a, b]$. For all C between $f(a)$ and $f(b)$ there exists a $c \in [a, b]$ such that $f(c) = C$. In other words, all intermediate values of a continuous function are obtained.*

Sketch of the proof. We proceed by **Divide and Conquer**. Without loss of generality, assume $f(a) < C < f(b)$. Let x_1 be the midpoint of $[a, b]$. If $f(x_1) = C$ we are done. If $f(x_1) < C$, we look at the interval $[x_1, b]$. If $f(x_1) > C$ we look at the interval $[a, x_1]$.

In either case, we have a new interval, call it $[a_1, b_1]$, such that $f(a_1) < C < f(b_1)$ and the interval has half the size of $[a, b]$. We continue in this manner, repeatedly taking the midpoint and looking at the appropriate half-interval.

If any of the midpoints satisfy $f(x_n) = C$, we are done. If no midpoint works, we divide infinitely often and obtain a sequence of points x_n in intervals $[a_n, b_n]$. This is where rigorous mathematical analysis is required (see §A.3 for a brief review, and [Rud] for complete details) to show x_n converges to an $x \in (a, b)$.

For each n we have $f(a_n) < C < f(b_n)$, and $\lim_{n\to\infty} |b_n - a_n| = 0$. As f is continuous, this implies $\lim_{n\to\infty} f(a_n) = \lim_{n\to\infty} f(b_n) = f(x) = C$. □

A.2.2 Mean Value Theorem

Theorem A.2.2 (Mean Value Theorem (MVT)). *Let $f(x)$ be differentiable on $[a, b]$. Then there exists a $c \in (a, b)$ such that*

$$f(b) - f(a) = f'(c) \cdot (b - a). \tag{A.14}$$

We give an interpretation of the Mean Value Theorem. Let $f(x)$ represent the distance from the starting point at time x. The average speed from a to b is the distance traveled, $f(b) - f(a)$, divided by the elapsed time, $b - a$. As $f'(x)$ represents the speed at time x, the Mean Value Theorem says that there is some intermediate time at which we are traveling at the average speed.

To prove the Mean Value Theorem, it suffices to consider the special case when $f(a) = f(b) = 0$; this case is known as Rolle's Theorem:

Theorem A.2.3 (Rolle's Theorem). *Let f be differentiable on $[a, b]$, and assume $f(a) = f(b) = 0$. Then there exists a $c \in (a, b)$ such that $f'(c) = 0$.*

Exercise A.2.4. *Show the Mean Value Theorem follows from Rolle's Theorem.* Hint: *Consider*

$$h(x) \ = \ f(x) - \frac{f(b) - f(a)}{b - a}(x - a) - f(a). \tag{A.15}$$

Note $h(a) = f(a) - f(a) = 0$ and $h(b) = f(b) - (f(b) - f(a)) - f(a) = 0$. The conditions of Rolle's Theorem are satisfied for $h(x)$, and

$$h'(c) \ = \ f'(c) - \frac{f(b) - f(a)}{b - a}. \tag{A.16}$$

Proof of Rolle's Theorem. Without loss of generality, assume $f'(a)$ and $f'(b)$ are non-zero. If either were zero we would be done. Multiplying $f(x)$ by -1 if needed, we may assume $f'(a) > 0$. *For convenience, we assume $f'(x)$ is continuous.* This assumption simplifies the proof, but is not necessary. In all applications in this book this assumption will be met.

Case 1: $f'(b) < 0$**:** As $f'(a) > 0$ and $f'(b) < 0$, the Intermediate Value Theorem applied to $f'(x)$ asserts that all intermediate values are attained. As $f'(b) < 0 < f'(a)$, this implies the existence of a $c \in (a, b)$ such that $f'(c) = 0$.

Case 2: $f'(b) > 0$**:** $f(a) = f(b) = 0$, and the function f is increasing at a and b. If x is real close to a then $f(x) > 0$ if $x > a$. This follows from the fact that

$$f'(a) \;=\; \lim_{x\to a} \frac{f(x) - f(a)}{x - a}. \tag{A.17}$$

As $f'(a) > 0$, the limit is positive. As the denominator is positive for $x > a$, the numerator must be positive. Thus $f(x)$ must be greater than $f(a)$ for such x. Similarly $f'(b) > 0$ implies $f(x) < f(b) = 0$ for x slightly less than b.

Therefore the function $f(x)$ is positive for x slightly greater than a and negative for x slightly less than b. If the first derivative were always positive then $f(x)$ could never be negative as it starts at 0 at a. This can be seen by again using the limit definition of the first derivative to show that if $f'(x) > 0$ then the function is increasing near x. Thus the first derivative cannot always be positive. Either there must be some point $y \in (a, b)$ such that $f'(y) = 0$ (and we are then done) or $f'(y) < 0$. By the Intermediate Value Theorem, as 0 is between $f'(a)$ (which is positive) and $f'(y)$ (which is negative), there is some $c \in (a, y) \subset [a, b]$ such that $f'(c) = 0$. □

A.2.3 Taylor Series

Using the Mean Value Theorem we prove a version of the n^{th} **Taylor series** Approximation: if f is differentiable at least $n+1$ times on $[a, b]$, then for all $x \in [a, b]$, $f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$ plus an error that is at most $\max_{a \le c \le x} |f^{(n+1)}(c)| \cdot |x-a|^{n+1}$.

Assuming f is differentiable $n+1$ times on $[a, b]$, we apply the Mean Value Theorem multiple times to bound the error between $f(x)$ and its Taylor Approximations. Let

$$\begin{aligned} f_n(x) \;&=\; \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \\ h(x) \;&=\; f(x) - f_n(x). \end{aligned} \tag{A.18}$$

$f_n(x)$ is the n^{th} Taylor series Approximation to $f(x)$. Note $f_n(x)$ is a polynomial of degree n and its first n derivatives agree with the derivatives of $f(x)$ at $x = 0$. We want to bound $|h(x)|$ for $x \in [a, b]$. Without loss of generality (basically, for notational convenience), we may assume $a = 0$. Thus $h(0) = 0$. Applying the Mean Value Theorem to h yields

$$\begin{aligned} h(x) &= h(x) - h(0) \\ &= h'(c_1) \cdot (x - 0) \quad \text{with } c_1 \in [0, x] \\ &= \left(f'(c_1) - f_n'(c_1)\right) x \\ &= \left(f'(c_1) - \sum_{k=1}^{n} \frac{f^{(k)}(0)}{k!} \cdot k(c_1 - 0)^{k-1}\right) x \\ &= \left(f'(c_1) - \sum_{k=1}^{n} \frac{f^{(k)}(0)}{(k-1)!} c_1^{k-1}\right) x \\ &= h_1(c_1)x. \end{aligned} \tag{A.19}$$

We now apply the Mean Value Theorem to $h_1(u)$. Note that $h_1(0) = 0$. Therefore

$$\begin{aligned} h_1(c_1) &= h_1(c_1) - h_1(0) \\ &= h_1'(c_2) \cdot (c_1 - 0) \quad \text{with } c_2 \in [0, c_1] \subset [0, x] \\ &= \left(f''(c_2) - f_n''(c_2)\right) c_1 \\ &= \left(f''(c_2) - \sum_{k=2}^{n} \frac{f^{(k)}(0)}{(k-1)!} \cdot (k-1)(c_2 - 0)^{k-2}\right) c_1 \\ &= \left(f''(c_2) - \sum_{k=2}^{n} \frac{f^{(k)}(0)}{(k-2)!} c_2^{k-2}\right) c_1 \\ &= h_2(c_2)c_1. \end{aligned} \tag{A.20}$$

Therefore,

$$h(x) \;=\; f(x) - f_n(x) \;=\; h_2(c_2)c_1x, \quad c_1, c_2 \in [0, x]. \tag{A.21}$$

Proceeding in this way a total of n times yields

$$h(x) \;=\; \left(f^{(n)}(c_n) - f^{(n)}(0)\right) c_{n-1}c_{n-2} \cdots c_2c_1x. \tag{A.22}$$

Applying the Mean Value Theorem to $f^{(n)}(c_n) - f^{(n)}(0)$ gives $f^{(n+1)}(c_{n+1}) \cdot (c_n - 0)$. Thus

$$h(x) \;=\; f(x) - f_n(x) \;=\; f^{(n+1)}(c_{n+1})c_n \cdots c_1x, \quad c_i \in [0, x]. \tag{A.23}$$

Therefore

$$|h(x)| \;=\; |f(x) - f_n(x)| \;\leq\; M_{n+1}|x|^{n+1} \tag{A.24}$$

where

$$M_{n+1} \;=\; \max_{c \in [0,x]} |f^{(n+1)}(c)|. \tag{A.25}$$

Thus if f is differentiable $n + 1$ times then the n^{th} Taylor series approximation to $f(x)$ is correct within a multiple of $|x|^{n+1}$; further, the multiple is bounded by the maximum value of $f^{(n+1)}$ on $[0, x]$.

Exercise A.2.5. *Prove* (A.22) *by induction.*

Exercise A.2.6. *Calculate the first few terms of the Taylor series expansions at* 0 *of* $\cos(x), \sin(x), e^x$*, and* $2x^3 - x + 3$*. Calculate the Taylor series expansions of the above functions at* $x = a$ Hint: *There is a fast way to do this.*

Exercise A.2.7 (Advanced)**.** *Show* all *the Taylor coefficients for*

$$f(x) \;=\; \begin{cases} e^{-1/x^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{A.26}$$

expanded about the origin vanish. What does this imply about the uniqueness of a Taylor series expansion? Warning: *be careful differentiating at zero. More is strangely true. Borel showed that if* $\{a_n\}$ *is any sequence of real numbers then there exists an infinitely differentiable* f *such that* $\forall n \geq 0$*,* $f^{(n)}(0) = a_n$ *(for a constructive proof see [GG]). Ponder the Taylor series from* $a_n = (n!)^2$*.*

A.2.4 Advanced Calculus Theorems

For the convenience of the reader we record exact statements of several standard results from advanced calculus that are used at various points of the text.

Theorem A.2.8 (Fubini)**.** *Assume* f *is continuous and*

$$\int_a^b \int_c^d |f(x,y)| dxdy \;<\; \infty. \tag{A.27}$$

Then

$$\int_a^b \left[\int_c^d f(x,y) dy \right] dx \;=\; \int_c^d \left[\int_a^b f(x,y) dx \right] dy. \tag{A.28}$$

Similar statements hold if we instead have

$$\sum_{n=N_0}^{N_1} \int_c^d f(x_n, y) dy, \qquad \sum_{n=N_0}^{N_1} \sum_{m=M_0}^{M_1} f(x_n, y_m). \tag{A.29}$$

For a proof in special cases, see [BL, VG]; an advanced, complete proof is given in [Fol]. See Exercise **??** for an example where the orders of integration cannot be changed.

Theorem A.2.9 (Green's Theorem)**.** *Let* C *be a simply closed, piecewise-smooth curve in the plane, oriented clockwise, bounding a region* D*. If* $P(x,y)$ *and* $Q(x,y)$ *have continuous partial derivatives on some open set containing* D*, then*

$$\int_C P(x,y)dx + Q(x,y)dy \;=\; \int\int_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dxdy. \tag{A.30}$$

For a proof, see [Rud], Theorem 9.50 as well as [BL, La5, VG].

Exercise A.2.10. *Prove Green's Theorem. Show it is enough to prove the theorem for* D *a rectangle, which is readily checked.*

Theorem A.2.11 (Change of Variables)**.** *Let V and W be bounded open sets in $\mathbb{R}^n$. Let $h : V \to W$ be a 1-1 and onto map, given by*

$$h(u_1, \dots, u_n) \;=\; (h_1(u_1, \dots, u_n), \dots, h_n(u_1, \dots, u_n))\,. \tag{A.31}$$

Let $f : W \to \mathbb{R}$ be a continuous, bounded function. Then

$$\begin{aligned}\int \cdots \int_W f(x_1, \dots, x_n) dx_1 \cdots dx_n \\ = \int \cdots \int_V f\left(h(u_1, \dots, u_n)\right) J(u_1, \dots, u_v) du_1 \cdots du_n.\end{aligned} \tag{A.32}$$

where J is the ***Jacobian***

$$J \;=\; \begin{vmatrix} \frac{\partial h_1}{\partial u_1} & \cdots & \frac{\partial h_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_n}{\partial u_1} & \cdots & \frac{\partial h_n}{\partial u_n} \end{vmatrix}. \tag{A.33}$$

For a proof, see [La5, Rud].

A.3 CONVERGENCE AND CONTINUITY

We recall some needed definitions and results from real analysis. See [Rud] for more details.

Definition A.3.1 (Convergence)**.** *A sequence $\{x_n\}_{n=1}^{\infty}$ converges to x if given any $\epsilon > 0$ there exists an N (possibly depending on ϵ) such that for all $n > N$, $|x_n - x| < \epsilon$. We often write $x_n \to x$.*

Exercise A.3.2. *If $x_n = \frac{3n^2}{n^2+1}$, prove $x_n \to 3$.*

Exercise A.3.3. *If $\{x_n\}$ converges, show it converges to a unique number.*

Exercise A.3.4. *Let $\alpha > 0$ and set $x_{n+1} = \frac{1}{2}\left(x_n + \frac{\alpha}{x_n}\right)$. If $x_0 = \alpha$, prove x_n converges to $\sqrt{\alpha}$. Can you generalize this to find p^{th} roots? This formula can be derived by Newton's Method (see §**??**).*

Definition A.3.5 (Continuity)**.** *A function f is continuous at a point x_0 if given an $\epsilon > 0$ there exists a $\delta > 0$ (possibly depending on ϵ) such that if $|x - x_0| < \delta$ then $|f(x) - f(x_0)| < \epsilon$.*

Definition A.3.6 (Uniform Continuity)**.** *A continuous function is uniformly continuous if given an $\epsilon > 0$ there exists a $\delta > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Note that the same δ works for all x.*

Usually we will work with functions that are uniformly continuous on some fixed, finite interval.

Theorem A.3.7. *Any continuous function on a closed, finite interval is uniformly continuous.*

Exercise A.3.8. *Show x^2 is uniformly continuous on $[a, b]$ for $-\infty < a < b < \infty$. Show $\frac{1}{x}$ is not uniformly continuous on $(0, 1)$, even though it is continuous. Show x^2 is not uniformly continuous on $[0, \infty)$.*

Exercise A.3.9. *Show the sum or product of two uniformly continuous functions is uniformly continuous. In particular, show any finite polynomial is uniformly continuous on $[a, b]$.*

We sketch a proof of Theorem A.3.7. We first prove

Theorem A.3.10 (Bolzano-Weierstrass)**.** *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in a finite closed interval. Then there is a subsequence $\{x_{n_k}\}_{k=1}^{\infty}$ such that x_{n_k} converges.*

Sketch the proof. Without loss of generality, assume the finite closed interval is $[0, 1]$. We proceed by divide and conquer. Consider the two intervals $I_1 = [0, \frac{1}{2}]$ and $I_2 = [\frac{1}{2}, 1]$. At least one of these (possibly both) must have infinitely many points of the original sequence as otherwise there would only be finitely many x_n's in the original sequence. Choose a subinterval (say I_a) with infinitely many x_n's, and choose any element of the sequence in that interval to be x_{n_1}.

Consider all x_n with $n > n_1$. Divide I_a into two subintervals I_{a1} and I_{a2} as before (each will be half the length of I_a). Again, at least one subinterval must contain infinitely many terms of the original sequence. Choose such a subinterval, say I_{ab}, and choose any element of the sequence in that interval to be x_{n_2} (note $n_2 > n_1$). We continue in this manner, obtaining a sequence $\{x_{n_k}\}$. For $k \geq K$, x_{n_k} is in an interval of size $\frac{1}{2^K}$. We we leave it as an exercise to the reader to show how this implies there is an x such that $x_{n_k} \to x$. □

Proof of Theorem A.3.7. If $f(x)$ is not uniformly continuous, given $\epsilon > 0$ for each $\delta = \frac{1}{2^n}$ there exist points x_n and y_n with $|x_n - y_n| < \frac{1}{2^n}$ and $|f(x_n) - f(y_n)| > \epsilon$. By the Bolzano-Weierstrass Theorem, we construct sequences $x_{n_k} \to x$ and $y_{n_{k_j}} \to y$. One can show $x = y$, and $|f(x_{n_{k_j}}) - f(y_{n_{k_j}})| > \epsilon$ violates the continuity of f at x. □

Exercise A.3.11. *Fill in the details of the above proof.*

Definition A.3.12 (Bounded)**.** *We say $f(x)$ is bounded (by B) if for all x in the domain of f, $|f(x)| \leq B$.*

Theorem A.3.13. *Let $f(x)$ be uniformly continuous on $[a, b]$. Then $f(x)$ is bounded.*

Exercise A.3.14. *Prove the above theorem. Hint: Given $\epsilon > 0$, divide $[a, b]$ into intervals of length δ.*

A.4 DIRICHLET'S PIGEON-HOLE PRINCIPLE

Theorem A.4.1 (Dirichlet's Pigeon-Hole Principle)**.** *Let $A_1, A_2, \dots, A_n$ be a collection of sets with the property that $A_1 \cup \cdots \cup A_n$ has at least $n + 1$ elements. Then at least one of the sets A_i has at least two elements.*

This is called the Pigeon-Hole Principle for the following reason: if $n+1$ pigeons go to n holes, at least one of the holes must be occupied by at least two pigeons. Equivalently, if we distribute k objects in n boxes and $k > n$, one of the boxes contains at least two objects. The Pigeon-Hole Principle is also known as the Box Principle. One application of the Pigeon-Hole Principle is to find good rational approximations to irrational numbers (see Theorem **??**). We give some examples to illustrate the method.

Example A.4.2. *If we choose a subset S from the set $\{1, 2, \dots, 2n\}$ with $|S| = n + 1$, then S contains at least two elements a, b with $a|b$.*

Write each element $s \in S$ as $s = 2^\sigma s_0$ with s_0 odd. There are n odd numbers in the set $\{1, 2, \dots, 2n\}$, and as the set S has $n + 1$ elements, the Pigeon-Hole Principle implies that there are at least two elements a, b with the same odd part; the result is now immediate.

Exercise A.4.3. *If we choose 55 numbers from $\{1, 2, 3, \dots, 100\}$ then among the chosen numbers there are two whose difference is ten (from [Ma]).*

Exercise A.4.4. *Let $a_1, \dots, a_{n+1}$ be distinct integers in $\{1, \dots, 2n\}$. Prove two of them add to a number divisible by $2n$.*

Exercise A.4.5. *Let $a_1, \dots, a_n$ be integers. Prove that there is a subset whose sum is divisible by n.*

Example A.4.6. *Let $\{a_1, a_2, a_3, a_4, a_5\}$ be distinct real numbers. There are indices i, j with $0 < a_i - a_j < 1 + a_i a_j$.*

As the function $\tan : (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$ is surjective, there are angles $\theta_i \in (-\frac{\pi}{2}, \frac{\pi}{2})$ with $a_i = \tan\theta_i$, $1 \le i \le 5$. Divide the interval $(-\frac{\pi}{2}, \frac{\pi}{2})$ into four equal pieces, each of length $\frac{\pi}{4}$. As we have five angles, at least two of them must lie in the same small interval, implying that there are i, j with $0 < \theta_i - \theta_j < \frac{\pi}{4}$. Applying $\tan$ to the last inequality and using the identity

$$\tan(x - y) \ = \ \frac{\tan x - \tan y}{1 + \tan x \tan y} \tag{A.34}$$

gives the result.

Exercise A.4.7. *Let $\phi_1, \phi_2, \dots, \phi_K$ be angles. Then for any $\epsilon > 0$ there are infinitely many $n \in \mathbb{N}$ such that*

$$\left| K - \sum_{j=1}^{K} \cos(n\phi_k) \right| \ < \ \epsilon. \tag{A.35}$$

Exercise$^{\text{(h)}}$ A.4.8. *The Pigeon-Hole Principle ensures that, if there are N boxes and $N + 1$ objects, then at least one box has two objects. What if we lower our sites and ask only that there is a high probability of having a box with two elements; see for example the birthday problem (Exercise 1.1.34). Specifically, let us assume that each object is equally likely to be in any of the N boxes. For each fixed k, show there is a positive probability of having at least k objects in a box if there are $N^{\frac{k-1}{k}}$ objects.*

A.5 MEASURES AND LENGTH

We discuss sizes of subsets of $[0, 1]$. It is natural to define the length of an interval $I = [a, b]$ (or $[a, b)$ and so on) as $b - a$. We denote this by $|I|$, and refer to this as the **length** or **measure** of I. Our definition implies a point a has zero length. What about more exotic sets, such as the rationals and the irrationals? What are the measures of these sets? A proper explanation is given by measure theory (see [La5, Rud]); we introduce enough for our purposes. We assume the reader is familiar with countable sets (see Chapter **??**).

Let I be a countable union of disjoint intervals $I_n \subset [0, 1)$; thus $I_n \cap I_m$ is empty if $n \neq m$. ***It is natural*** (but see §**??** as a warning for how *natural* statements are often wrong) to say

$$|I| \;=\; \sum_n |I_n|. \tag{A.36}$$

It is important to take a countable union. Consider an uncountable union with $I_x = \{x\}$ for $x \in [0, 1]$. As each singleton $\{x\}$ has length zero, we expect their union to also have length zero; however, their union is $[0, 1]$, which has length 1. If $A \subset B$, it is natural to say $|A|$ (the length of A) is at most $|B|$ (the length of B). Note our definition implies $[a, b)$ and $[a, b]$ have the same length.

A.5.1 Measure of the Rationals

Our assumptions imply that the rationals in $[0, 1]$ have zero length (hence the irrationals in $[0, 1]$ have length 1).

Theorem A.5.1. *The rationals $\mathbb{Q}$ have zero measure.*

Sketch of the proof. We claim it suffices to show $Q = \mathbb{Q} \cap [0, 1]$ has measure zero. To prove $|Q| = 0$ we show that given any $\epsilon > 0$ we can find a countable set of intervals I_n such that

1. $|Q| \subset \cup_n I_n$;
2. $\sum_n |I_n| < \epsilon$.

As the rationals are countable, we can enumerate Q, say $Q = \{x_n\}_{n=0}^{\infty}$. For each n let

$$I_n \;=\; \left[x_n - \frac{\epsilon}{4 \cdot 2^n},\; x_n + \frac{\epsilon}{4 \cdot 2^n}\right], \quad |I_n| \;=\; \frac{\epsilon}{2 \cdot 2^n}. \tag{A.37}$$

Clearly $Q \subset \cup_n I_n$. The intervals I_n are not necessarily disjoint, but

$$|\cup_n I_n| \;\leq\; \sum_n |I_n| \;=\; \epsilon, \tag{A.38}$$

which completes the proof. □

Exercise A.5.2. *Show that if $Q = \mathbb{Q} \cap [0, 1]$ has measure zero, then $\mathbb{Q}$ has measure zero.*

Exercise A.5.3. *Show any countable set has measure zero; in particular, the algebraic numbers have length zero.*

Definition A.5.4 (Almost all)**.** *Let A^c be the compliment of $A \subset \mathbb{R}$: $A^c = \{x : x \notin A\}$. If A^c is of measure zero, we say almost all x are in A.*

Thus the above theorem shows that not only are almost all real numbers are irrational but almost all real numbers are transcendental.

A.5.2 Measure of the Cantor Set

The Cantor set is a fascinating subset of $[0,1]$. We construct it in stages. Let $C_0 = [0,1]$. We remove the middle third of C_0 and obtain $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Note C_1 is a union of two closed intervals (we keep all endpoints). To construct C_2 we remove the middle third of all remaining intervals and obtain

$$C_2 \ = \ \left[0, \ \frac{1}{9}\right] \bigcup \left[\frac{2}{9}, \ \frac{3}{9}\right] \bigcup \left[\frac{6}{9}, \ \frac{7}{9}\right] \bigcup \left[\frac{8}{9}, \ 1\right]. \tag{A.39}$$

We continue this process. Note C_n is the union of 2^n closed intervals, each of size 3^{-n}, and

$$C_0 \ \supset \ C_1 \ \supset \ C_2 \ \supset \ \cdots. \tag{A.40}$$

Definition A.5.5 (Cantor Set)**.** *The Cantor set C is defined by*

$$C \ = \ \bigcap_{n=1}^{\infty} C_n \ = \ \{x \in \mathbb{R} : \forall n, x \in C_n\}. \tag{A.41}$$

Exercise A.5.6. *Show the length of the Cantor set is zero.*

If x is an endpoint of C_n for some n, then $x \in C$. At first, one might expect that these are the only points, especially as the Cantor set has length zero.

Exercise A.5.7. *Show $\frac{1}{4}$ and $\frac{3}{4}$ are in C, but neither is an endpoint.* Hint: *Proceed by induction. To construct C_{n+1} from C_n, we removed the middle third of intervals. For each sub-interval, what is left looks like the union of two pieces, each one-third the length of the previous. Thus, we have shrinking maps fixing the left and right parts $L, R : \mathbb{R} \to \mathbb{R}$ given by $L(x) = \frac{x}{3}$ and $R(x) = \frac{x+2}{3}$, and $C_{n+1} = R(C_n) + L(C_n)$.*

Exercise A.5.8. *Show the Cantor set is also the set of all numbers $x \in [0,1]$ which have no 1's in their base three expansion. For rationals such as $\frac{1}{3}$, we may write these by using repeating 2's: $\frac{1}{3} = .02222\ldots$ in base three. By considering base two expansions, show there is a one-to-one and onto map from $[0,1]$ to the Cantor set.*

Exercise A.5.9 (From the *American Mathematical Monthly*)**.** *Use the previous exercise to show that every $x \in [0,2]$ can be written as a sum $y + z$ with $y, z \in C$.*

Remark A.5.10. The above exercises show the Cantor set is uncountable and is in a simple correspondence to all of $[0, 1]$, *but* it has length zero! Thus, the notion of "length" is different from the notion of "cardinality": two sets can have the same cardinality but very different lengths.

Exercise A.5.11 (Fat Cantor Sets)**.** *Instead of removing the middle third in each step, remove the middle $\frac{1}{m}$. Is there a choice of m which yields a set of positive length? What if at stage n we remove the middle $\frac{1}{a_n}$. For what sequences a_n are we left with a set of positive length? If the a_n are digits of a simple continued fraction, what do you expect to be true for "most" such numbers?*

For more on the Cantor set, including dynamical interpretations, see [Dev, Edg, Fal, SS3].

A.6 INEQUALITIES

The first inequality we mention here is the Arithmetic Mean and Geometrically Mean Inequality (AM–GM); see [Mil3] for some proofs. For positive numbers $a_1, \dots, a_n$, the arithmetic mean is $\frac{a_1+\cdots+a_n}{n}$ and the geometric mean is $\sqrt[n]{a_1 \cdots a_n}$.

Theorem A.6.1 (AM-GM)**.** *Let $a_1, \dots, a_n$ be positive real numbers. Then*

$$\sqrt[n]{a_1 \cdots a_n} \ \le \ \frac{a_1 + \cdots + a_n}{n}, \tag{A.42}$$

with equality if and only if $a_1 = \cdots = a_n$.

Exercise A.6.2. *Prove the AM-GM when $n = 2$.* Hint: *For $x \in \mathbb{R}$, $x^2 \ge 0$; this is one of the most useful inequalities in mathematics. We will see it again when we prove the Cauchy-Schwartz inequality.*

Exercise A.6.3. *Prove the AM-GM using mathematical induction.*

There is an interesting generalization of the AM-GM; AM-GM is the case $p_1 = \cdots = p_n = \frac{1}{n}$ of the following theorem.

Theorem A.6.4. *Let $a_1, \dots, a_n$ be as above, and let $p_1, \dots, p_n$ be positive real numbers. Set $P = p_1 + \cdots + p_n$. Then*

$$a_1^{p_1} \dots a_n^{p_n} \ \le \ \left(\frac{p_1 a_1 + \cdots + p_n a_n}{P}\right)^P, \tag{A.43}$$

and equality holds if and only if $a_1 = \cdots = a_n$.

This inequality is in turn a special case of the following important theorem:

Theorem A.6.5 (Jensen's Inequality)**.** *Let f be a real continuous function on $[a, b]$ with continuous second derivative on (a, b). Suppose that $f''(x) \le 0$ for all $x \in (a, b)$. Then for $a_1, \dots, a_n \in [a, b]$ and $p_1, \dots, p_n$ positive real numbers, we have*

$$f\left(\frac{p_1 a_1 + \cdots + p_n a_n}{p_1 + \cdots + p_n}\right) \ \le \ \frac{p_1 f(a_1) + \cdots + p_n f(a_n)}{p_1 + \cdots + p_n}. \tag{A.44}$$

Exercise A.6.6. *Prove Jensen's inequality.* Hint: *Draw a picture; carefully examine the case $n = 2$, $p_1 = p_2 = \frac{1}{2}$. What does $f''(x) \leq 0$ mean in geometric terms?*

Exercise A.6.7. *Investigate the cases where Jensen's inequality is an equality.*

Exercise A.6.8. *Show that Jensen's inequality implies the AM-GM and its generalization Theorem A.6.4.* Hint: *Examine the function $f(x) = -\log x$, $x > 0$.*

Our final inequality is the **Cauchy-Schwarz inequality**. There are a number of inequalities that are referred to as the Cauchy-Schwarz inequality. A useful version is the following:

Lemma A.6.9 (Cauchy-Schwarz)**.** *For complex-valued functions f and g,*

$$\int_0^1 |f(x)g(x)|dx \ \leq \ \left(\int_0^1 |f(x)|^2 dx\right)^{\frac{1}{2}} \cdot \left(\int_0^1 |g(x)|^2 dx\right)^{\frac{1}{2}}. \qquad \text{(A.45)}$$

Proof. For notational simplicity, assume f and g are non-negative functions. Working with $|f|$ and $|g|$ we see there is no harm in the above assumption. As the proof is immediate if either of the integrals on the right hand side of (A.45) is zero or infinity, we assume both integrals are non-zero and finite. Let

$$h(x) = \ f(x) - \lambda g(x), \ \ \lambda \ = \ \frac{\int_0^1 f(x)g(x)dx}{\int_0^1 g(x)^2 dx}. \qquad \text{(A.46)}$$

As $\int_0^1 h(x)^2 dx \geq 0$ we have

$$\begin{aligned} 0 \ &\leq \ \int_0^1 \left(f(x) - \lambda g(x)\right)^2 dx \\ &= \ \int_0^1 f(x)^2 dx \ - \ 2\lambda \int_0^1 f(x)g(x)dx \ + \ \lambda^2 \int_0^1 g(x)^2 dx \\ &= \ \int_0^1 f(x)^2 dx \ - \ 2\frac{\left(\int_0^1 f(x)g(x)dx\right)^2}{\int_0^1 g(x)^2 dx} \ + \ \frac{\left(\int_0^1 f(x)g(x)dx\right)^2}{\int_0^1 g(x)^2 dx} \\ &= \ \int_0^1 f(x)^2 dx \ - \ \frac{\left(\int_0^1 f(x)g(x)dx\right)^2}{\int_0^1 g(x)^2 dx}. \qquad \text{(A.47)} \end{aligned}$$

This implies

$$\frac{\left(\int_0^1 f(x)g(x)dx\right)^2}{\int_0^1 g(x)^2 dx} \ \leq \ \int_0^1 f(x)^2 dx, \qquad \text{(A.48)}$$

or equivalently

$$\left(\int_0^1 f(x)g(x)dx\right)^2 \ \leq \ \int_0^1 f(x)^2 dx \cdot \int_0^1 g(x)^2 dx. \qquad \text{(A.49)}$$

Taking square roots completes the proof. □

Again, note that both the AG-GM and the Cauchy-Schwartz inequalities are clever applications of $x^2 \geq 0$ for $x \in \mathbb{R}$.

Exercise A.6.10. *For what f and g is the Cauchy-Schwarz Inequality an equality?*

Exercise A.6.11. *One can also prove the Cauchy-Schwartz inequality as follows: consider $h(x) = af(x) + bg(x)$ where $a = \sqrt{\int_0^1 |f(x)|^2 dx}$, $b = \sqrt{\int_0^1 |g(x)|^2 dx}$, and then integrate $h(x)^2$.*

Remark A.6.12. The Cauchy-Schwarz Inequality is often useful when $g(x) = 1$. In this special case, it is important that we integrate over a finite interval.

Exercise A.6.13. *Suppose $a_1, \dots, a_n$ and $b_1, \dots, b_n$ are two sequences of real numbers. Prove the following Cauchy-Schwarz inequality:*

$$|a_1 b_1 + a_2 b_2 + \cdots + a_n b_n| \ \leq \ (a_1^2 + \dots a_n^2)^{\frac{1}{2}} (b_1^2 + \cdots + b_n^2)^{\frac{1}{2}}. \qquad \text{(A.50)}$$

Exercise A.6.14. *Let $f, g : \mathbb{R} \to \mathbb{C}$ be such that $\int_{\mathbb{R}} |f(x)|^2 dx, \int_{\mathbb{R}} |g(x)|^2 dx < \infty$. Prove the following Cauchy-Schwarz inequality:*

$$\left| \int_{-\infty}^{\infty} f(x) g(x) dx \right|^2 \ \leq \ \int_{-\infty}^{\infty} |f(x)|^2 dx \cdot \int_{-\infty}^{\infty} |g(x)|^2 dx. \qquad \text{(A.51)}$$

Appendix B

Hints and Remarks on the Exercises

Chapter 8: Introduction to Probability

Exercise 1.1.18: *Hint:* Let a_n be the probability that there are at least 3 consecutive heads in n tosses. Show a_n satisfies the recurrence relation

$$a_n \;=\; \frac{1}{2}a_{n-1} + \frac{1}{4}a_{n-2} + \frac{1}{8}a_{n-3} + \frac{1}{8}. \tag{B.1}$$

The presence of the final term, $\frac{1}{8}$, greatly complicates matters; we cannot use the methods of Exercise **??** or §2.3 to solve the recurrence relation. It is much easier to study b_n, the probability that there are not 3 consecutive heads in n tosses; note $a_n = 1 - b_n$. Show b_n satisfies

$$b_n \;=\; \frac{1}{2}b_{n-1} + \frac{1}{4}b_{n-2} + \frac{1}{8}b_{n-3}. \tag{B.2}$$

More generally, determine the probability of observing at least k heads in n tosses of a coin with probability p of heads. If $p = \frac{1}{2}$ show that the roots of the characteristic polynomial of the recurrence relation are at most $\left(1 - 2^{-k}\right)^{1/k}$. One application of this is to roulette, where the probability of getting red (or black) is $16/38$ because there are two green spaces. This shows there is a large enough probability of consecutive losses so that the strategy of double plus one (bet \$1 on the first spin; if you lose bet \$2 on the second, if you lose again bet \$4 on the third, if you lose again bet \$8 on the fourth, and so on; it does not matter when your color finally comes up – you always win \$1) will fail in general, as too quickly you reach the house limit (maximum allowable bet) and lose a lot.

Exercise 1.1.36: *Hint:* Let $X_{[m]}$ denote the largest of player one's rolls, and $Y_{[n]}$ the largest of player two's rolls. For $a \in \{1, \dots, k\}$,

$$\text{Prob}(X_{[m]} = a) \;=\; \frac{a^m - (a-1)^m}{k^m}; \tag{B.3}$$

this follows from

$$\text{Prob}(X_{[m]} = a) \;=\; \sum_{\ell=1}^{m} \binom{m}{\ell} \frac{1}{k^\ell} \left(\frac{a-1}{k}\right)^{m-\ell}, \tag{B.4}$$

the binomial theorem and noticing we have a telescoping sum. The proof is completed by noting that

$$\text{Prob(Player one wins)} \;=\; \sum_{a=2}^{k} \text{Prob}(X_{[m]} = a) \cdot \text{Prob}(Y_{[n]} \le a-1). \tag{B.5}$$

$X_{[m]}$ and $Y_{[n]}$ are examples of order statistics; see also Exercise **??**.

Exercise 1.2.9: *Hint:* Let

$$f(\lambda) \;=\; \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \;=\; e^{\lambda}. \tag{B.6}$$

Differentiate once to determine the mean, twice to determine the variance.

Chapter 9: Applications of Probability: Benford's Law and Hypothesis Testing

Exercise 2.3.3: *Hint:* Consider $a_0 = a_1 = a_2 = 1$. This recurrence relation was constructed by starting with the characteristic polynomial $(r-2)^2(r-1)$ and then finding initial conditions so that the coefficients of the $\lambda_1 = \lambda_2 = 2$ eigenvalues vanish. In searching for counter-examples, it is significantly easier here to specify the roots of the characteristic polynomial first, and find the actual recurrence relation second.

Exercise 2.3.4: *Hint:* Consider a recurrence relation of length k with k distinct roots. By specifying k terms (say $a_0, \dots, a_{k-1}$), the coefficients of the roots λ_i are determined. We must solve

$$u_1\lambda_1^n + \cdots + u_k\lambda_k^n \;=\; a_n, \quad n \in \{0, \dots, k-1\}. \tag{B.7}$$

We may write this in matrix form as

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_k^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{k-1} & \lambda_2^{k-1} & \cdots & \lambda_k^{k-1} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_k \end{pmatrix} \;=\; \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{k-1} \end{pmatrix}. \tag{B.8}$$

The matrix of eigenvalues is a Vandermonde matrix; by Exercise **??** its determinant is non-zero when $\lambda_i \neq \lambda_j$. Thus its inverse exists, and the initial conditions which lead to $u_1 = 0$ are a hyperplane in $\mathbb{C}^k$, which shows that almost all initial conditions lead to $u_1 \neq 0$.

Appendix C

Concluding Remarks

This book is meant as an introduction to a vast, active subject. It is our hope that the reader will pursue these topics further through the various projects and references mentioned in the introduction and chapters above. We also hope that we have shown how similar tools, techniques and concepts arise in different parts of mathematics. We briefly summarize some of what we have seen.

The first is the Philosophy of Square Root Cancellation. As a general principle, many “nice” sums of N terms of absolute value 1 are approximately of size $\sqrt{N}$. Examples range from the Gauss sums of §**??** (which were then used in our investigations of the number of solutions to Diophantine equations in §**??**) to the average value of generating functions encountered in the Circle Method in Chapters **??** and **??** to the Central Limit Theorem of §1.4 (which shows that for a wide class of populations, the distribution of the mean of a large sample is independent of the fine properties of the underlying distribution).

Similar to the universality of the Central Limit Theorem, many different systems after normalization follow the same spacing laws. We have seen numerical and theoretical evidence showing that spacings between primes, the fractional parts of $n^k\alpha$ (for certain k and α) and numbers uniformly chosen in $[0, 1]$ are the same (see Chapter **??**), while in Chapters **??** to **??** we see similar behavior in energy levels of heavy nuclei, eigenvalues of matrices (of random matrix ensembles as well as adjacency matrices attached to d-regular graphs) and zeros of L-functions.

Throughout our investigations, certain viewpoints have consistently proven useful. Among the most important are Fourier Analysis (Chapter **??**) and the structure of numbers (Chapters **??** and **??**). From Fourier Analysis we obtain Poisson Summation and the Fourier Transform (which are useful for investigating problems as varied as the first digits of sequences (§2.4.2), the functional equation of $\zeta(s)$ (§**??**) and in Chapter **??** the zeros of L-functions). Other applications range from Weyl’s Theorem (Chapter **??**) on the equidistribution of sequences to the Circle Method and representing numbers as the sum of primes or integer powers (§**??** and **??**). We have used the structure of numbers in finding good rational approximations (§**??**), Roth’s Theorem (Chapter **??**), and studying the properties of $n^k\alpha \bmod 1$ (Chapter **??**).

Finally, we have tried to emphasize in the text which techniques appear throughout mathematics. Some of the most common are adding zero or multiplying by one, divide and conquer, dyadic decomposition, no integers are in $(0, 1)$, the Pigeon-Hole Principle, positivity, and splitting integrals or sums; see the *techniques* entry in the index for more details.

Bibliography

Links to many of the references below are available online at
http://www.math.princeton.edu/mathlab/book/index.html

[Acz] A. Aczel, *Fermat's Last Theorem: Unlocking the Secret of an Ancient Mathematical Problem*, Four Walls Eight Windows, New York, 1996.

[AKS] R. Adler, M. Keane, and M. Smorodinsky, *A construction of a normal number for the continued fraction transformation*, J. Number Theory **13** (1981), no. 1, 95–105.

[AgKaSa] M. Agrawal, N. Kayal and N. Saxena, *PRIMES is in P*, Ann. of Math. (2) **160** (2004), no. 2, 781–793.

[Al] L. Ahlfors, *Complex Analysis*, 3rd edition, McGraw-Hill, New York, 1979.

[AZ] M. Aigner and G. M. Ziegler, *Proofs from THE BOOK*, Springer-Verlag, Berlin, 1998.

[AGP] W. R. Alford, A. Granville, and C. Pomerance, *There are infinitely many Carmichael numbers*, Ann. Math. **139** (1994), 703–722.

[AMS] AMS MathSciNet, http://www.ams.org/msnmain?screen=Review

[AB] U. Andrews IV and J. Blatz, *Distribution of digits in the continued fraction representations of seventh degree algebraic irrationals*, Junior Thesis, Princeton University, Fall 2002.

[Ap] R. Apéry, *Irrationalité de $\zeta(2)$ et $\zeta(3)$*, Astérisque **61** (1979) 11–13.

[Apo] T. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, New York, 1998.

[ALM] S. Arms, A. Lozano-Robledo and S. J. Miller, *Constructing One-Parameter Families of Elliptic Curves over $\mathbb{Q}(T)$ with Moderate Rank*, to appear in the Journal of Number Theory.

[Art] M. Artin, *Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[Ay] R. Ayoub, *Introduction to the Analytic Theory of Numbers*, AMS, Providence, RI, 1963.

[Bai] Z. Bai, *Methodologies in spectral analysis of large-dimensional random matrices, a review*, Statist. Sinica **9** (1999), no. 3, 611–677.

[B] A. Baker, *Transcendental Number Theory*, Cambridge University Press, Cambridge, 1990.

[BM] R. Balasubramanian and C. J. Mozzochi, *Siegel zeros and the Goldbach problem*, J. Number Theory **16** (1983), no. 3, 311–332.

[BR] K. Ball and T. Rivoal, *Irrationalité d'une infinité valeurs de la fonction zeta aux entiers impairs*, Invent. Math. **146** (2001), 193–207.

[BT] V. V. Batyrev and Yu. Tschinkel, *Tamagawa numbers of polarized algebraic varieties*, Nombre et répartition de points de hauteur bornée (Paris, 1996), Astérisque (1998), No. 251, 299–340.

[BL] P. Baxandall and H. Liebeck, *Vector Calculus*, Clarendon Press, Oxford, 1986.

[Be] R. Beals, *Notes on Fourier series*, Lecture Notes, Yale University, 1994.

[Bec] M. Beceanu, *Period of the continued fraction of* $\sqrt{n}$, Junior Thesis, Princeton University, 2003.

[Ben] F. Benford, *The law of anomalous numbers*, Proceedings of the American Philosophical Society **78** (1938) 551–572.

[BBH] A. Berger, Leonid A. Bunimovich, and T. Hill, *One-dimensional dynamical systems and Benford's Law*, Trans. Amer. Math. Soc. **357** (2005), no. 1, 197–219.

[BEW] B. Berndt, R. Evans, and K. Williams, *Gauss and Jacobi Sums*, Canadian Mathematical Society Series of Monographs and Advanced Texts, Vol. 21, Wiley-Interscience Publications, John Wiley & Sons, New York, 1998.

[Ber] M. Bernstein, *Games, hats, and codes*, lecture at the SUMS 2005 Conference.

[BD] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, Holden-Day, San Francisco, 1977.

[Bi] P. Billingsley, *Probability and Measure*, 3rd edition, Wiley, New York, 1995.

[Bl1] P. Bleher, *The energy level spacing for two harmonic oscillators with golden mean ratio of frequencies*, J. Stat. Phys. **61** (1990) 869–876.

[Bl2] P. Bleher, *The energy level spacing for two harmonic oscillators with generic ratio of frequencies*, J. Stat. Phys. **63** (1991), 261–283.

[Bob] J. Bober, *On the randomness of modular inverse mappings*, Undergraduate Mathematics Laboratory report, Courant Institute, NYU, 2002.

[Bol] B. Bollobás, *Random Graphs*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge, 2001.

[BoLa] E. Bombieri and J. Lagarias, *Complements to Li's criterion for the Riemann Hypothesis*, J. Number Theory **77** (1999), no. 2, 274–287.

[BG] E. Bombieri and W. Gubler, *Heights in Diophantine Geometry*, Cambridge University Press, Cambridge, UK, 2006.

[BP] E. Bombieri and A. van der Poorten, *Continued fractions of algebraic numbers*. Pages 137–152 in *Computational Algebra and Number Theory (Sydney, 1992)*, Mathematical Applications, Vol. 325, Kluwer Academic, Dordrecht, 1995.

[Bon] D. Boneh, *Twenty years of attacks on the RSA cryptosystem*, Notices of the American Mathematical Society **46** (1999), no. 2, 203–213.

[BS] Z. Borevich and I. Shafarevich, *Number Theory*, Academic Press, New York, 1968.

[BB] J. Borwein and P. Borwein, *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*, John Wiley and Sons, New York, 1987.

[BK] A. Boutet de Monvel and A. Khorunzhy, *Some elementary results around the Wigner semicircle law*, lecture notes.

[BoDi] W. Boyce and R. DiPrima, *Elementary differential equations and boundary value problems*, 7th edition, John Wiley & Sons, New York, 2000.

[Bre1] R. Brent, *The distribution of small gaps between successive primes*, Math. Comp. **28** (1974), 315–324.

[Bre2] R. Brent, *Irregularities in the distribution of primes and twin primes*, Collection of articles dedicated to Derrick Henry Lehmer on the occasion of his seventieth birthday, Math. Comp. **29** (1975), 43–56.

[BPR] R. Brent, A. van der Poorten, and H. te Riele, *A comparative study of algorithms for computing continued fractions of algebraic numbers*. Pages 35–47 in *Algorithmic number theory (Talence, 1996)*, Lecture Notes in Computer Science, Vol. 1122, Springer, Berlin, 1996.

[deBr] R. de la Bretèche, *Sur le nombre de points de hauteur bornée d'une certaine surface cubique singulière*. Pages 51–77 in *Nombre et répartition de points de hauteur bornée (Paris, 1996)*, Astérisque, (1998) no. 251, 51–77.

[BBD] R. de la Bretèche, T. D. Browning, and U. Derenthal, *On Manin's conjecture for a certain singular cubic surface*, preprint.

[BPPW] B. Brindza, A. Pintér, A. van der Poorten, and M. Waldschmidt, *On the distribution of solutions of Thue's equation*. Pages 35–46 in *Number theory in progress (Zakopane-Koscielisko, 1997)*, Vol. 1, de Gruyter, Berlin, 1999.

[BFFMPW] T. Brody, J. Flores, J. French, P. Mello, A. Pandey, and S. Wong, *Random-matrix physics: spectrum and strength fluctuations*, Rev. Mod. Phys. **53** (1981), no. 3, 385–479.

[BrDu] J. Brown and R. Duncan, *Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences*, Fibonacci Quarterly **8** (1970) 482–486.

[Bro] T. Browning, *The density of rational points on a certain singular cubic surface*, preprint.

[BDJ] W. Bryc, A. Dembo, T. Jiang, *Spectral measure of large random Hankel, Markov and Toeplitz matrices*, Ann. Probab. **34** (2006), no. 1, 1–38.

[Bry] A. Bryuno, *Continued frations of some algebraic numbers*, U.S.S.R. Comput. Math. & Math. Phys. **4** (1972), 1–15.

[Bur] E. Burger, *Exploring the Number Jungle: A Journey into Diophantine Analysis*, AMS, Providence, RI, 2000.

[BuP] E. Burger and A. van der Poorten, *On periods of elements from real quadratic number fields*. Pages 35–43 in *Constructive, Experimental, and Nonlinear Analysis (Limoges, 1999)*, CMS Conf. Proc., **27**, AMS, Providence, RI, 2000.

[CaBe] G. Casella and R. Berger, *Statistical Inference*, 2nd edition, Duxbury Advanced Series, Pacific Grove, CA, 2002.

[CGI] G. Casati, I. Guarneri, and F. M. Izrailev, *Statistical properties of the quasi-energy spectrum of a simple integrable system*, Phys. Lett. A **124** (1987), 263–266.

[Car] L. Carleson, *On the convergence and growth of partial sums of Fourier series*, Acta Math. **116** (1966), 135–157.

[Ca] J. W. S. Cassels, *An Introduction to Diophantine Approximation*, Cambridge University Press, London 1957.

[Ch] D. Champernowne, *The construction of decimals normal in the scale of ten*, J. London Math. Soc. 8 (1933), 254–260.

[Cha] K. Chang, *An experimental approach to understanding Ramanujan graphs*, Junior Thesis, Princeton University, Spring 2001.

[ChWa] J. R. Chen and T. Z. Wang, *On the Goldbach problem*, Acta Math. Sinica **32** (1989), 702–718.

[Chr] J. Christiansen, *An introduction to the moment problem*, lecture notes.

[Ci] J. Cisneros, *Waring's problem*, Junior Thesis, Princeton University, Spring 2001.

[CW] J. Coates and A. Wiles, *On the conjecture of Birch and Swinnterton-Dyer*, Invent. Math. **39** (1977), 43–67.

[CB] S. Chatterjee and A. Bose, *A new method for bounding rates of convergence of empirical spectral distributions*, J. Theoret. Probab. **17** (2004), no. 4, 1003–1019.

[Cof1] M. Coffey, *Toward verification of the Riemann hypothesis: Application of the Li criterion*, to appear in Math. Physics, Analysis and Geometry. http://arxiv.org/pdf/math-ph/0505052.

[Cof2] M. Coffey, *On the coefficients of the Baez-Duarte criterion for the Riemann hypothesis and their extensions*, preprint. http://arxiv.org/pdf/math-ph/0608050.

[CL1] H. Cohen and H. W. Lenstra, Jr., *Heuristics on class groups of number fields*. Pages 33–62 in *Number Theory*, Lecture Notes in Mathematics, Vol. 1068, Springer-Verlag, Berlin, 33–62.

[CL2] H. Cohen and H. W. Lenstra, Jr., *Heuristics on class groups*, in *Number Theory*, Lecture Notes in Mathematics, Vol. 1052, Springer-Verlag, Berlin, 26–36.

[Coh] P. Cohen, *The independence of the continuum hypothesis*, Proc. Nat. Acad. Sci. U.S.A, **50** (1963), 1143–1148; **51** (1964), 105–110.

[Cohn] J. Cohn, *The length of the period of simple continued fractions*, Pacific Journal of Mathematics, **71** (1977), no. 1, 21–32.

[Con1] J. B. Conrey, *L-Functions and random matrices*. Pages 331–352 in *Mathematics unlimited — 2001 and Beyond*, Springer-Verlag, Berlin, 2001.

[Con2] J. B. Conrey, *The Riemann hypothesis*, Notices of the AMS, **50** (2003), no. 3, 341–353.

[CFKRS] B. Conrey, D. Farmer, P. Keating, M. Rubinstein and N. Snaith, *Integral moments of L-functions*, Proc. London Math. Soc. (3) **91** (2005), no. 1, 33–104.

[Conw] J. H. Conway, *The weird and wonderful chemistry of audioactive decay*. Pages 173–178 in *Open Problems in Communications and Computation*, ed. T. M. Cover and B. Gopinath, Springer-Verlag, New York, 1987.

[CG] J. H. Conway and R. Guy, *The Book of Numbers*, Springer-Verlag, Berlin, 1996.

[CS] J. H. Conway and N. J. A. Sloane, *Lexicographic Codes: Error-Correcting Codes from Game Theory*, IEEE Trans. Inform. Theory, **32** (1986), no. 3, 219–235.

[Corl] R. M. Corless,*Continued fractions and chaos*. Amer. Math. Monthly **99** (1992), no. 3, 203–215.

[Cor1] Cornell University, *arXiv*, http://arxiv.org

[Cor2] Cornell University, *Project Euclid*, http://projecteuclid.org/

[CFS] I. P. Cornfeld, S. V. Fomin, and I. G. Sinai, *Ergodic Theory*, Grundlehren Der Mathematischen Wissenschaften, Springer-Verlag, Berlin, 1982.

[Da1] H. Davenport, *The Higher Arithmetic: An Introduction to the Theory of Numbers*, 7th edition, Cambridge University Press, Cambridge, 1999.

[Da2] H. Davenport, *Multiplicative Number Theory*, 2nd edition, revised by H. Montgomery, Graduate Texts in Mathematics, Vol. 74, Springer-Verlag, New York, 1980.

[Da3] H. Davenport, *On the distribution of quadratic residues (mod p)*, London Math. Soc. **6** (1931), 49–54.

[Da4] H. Davenport, *On character sums in finite fields*, Acta Math. **71** (1939), 99–121.

[DN] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd edition, Wiley Interscience, Hoboken, NJ, 2003.

[DSV] G. Davidoff, P. Sarnak, and A. Valette, *Elementary Number Theory, Group Theory, and Ramanujan Graphs*, London Mathematical Society, Student Texts, Vol. 55, Cambridge University Press, Cambridge 2003.

[Dev] R. Devaney, *An Introduction to Chaotic Dynamical Systems*, 2nd edition, Westview Press, Cambridge, MA, 2003.

[Dia] P. Diaconis, *Patterns in eigenvalues: the* 70^{th} *Josiah Williard Gibbs lecture*, Bulletin of the American Mathematical Society **40** (2003), no. 2, 155–178.

[Di] T. Dimofte, *Rational shifts of linearly periodic continued fractions*, Junior Thesis, Princeton University, 2003.

[DM] E. Dueñez and S. J. Miller, *The Low Lying Zeros of a* $\mathrm{GL}(4)$ *and a* $\mathrm{GL}(6)$ *family of L-functions*, to appear in Compositio Mathematica.

[Du] R. Durrett, *Probability: Theory and Examples*, 2nd edition, Duxbury Press, 1996.

[Dy1] F. Dyson, *Statistical theory of the energy levels of complex systems: I, II, III*, J. Mathematical Phys. **3** (1962) 140–156, 157–165, 166–175.

[Dy2] F. Dyson, *The threefold way. Algebraic structure of symmetry groups and ensembles in quantum mechanics*, J. Mathematical Phys., **3** (1962) 1199–1215.

[Edg] G. Edgar, *Measure, Topology, and Fractal Geometry*, 2nd edition, Springer-Verlag, 1990.

[Ed] H. M. Edwards, *Riemann's Zeta Function*, Academic Press, New York, 1974.

[EST] B. Elias, L. Silberman and R. Takloo-Bighash, *On Cayley's theorem*, preprint.

[EE] W. J. Ellison and F. Ellison, *Prime Numbers*, John Wiley & Sons, New York, 1985.

[Est1] T. Estermann, *On Goldbach's problem: Proof that almost all even positive integers are sums of two primes*, Proc. London Math. Soc. Ser. 2 **44** (1938) 307–314.

[Est2] T. Estermann, *Introduction to Modern Prime Number Theory*, Cambridge University Press, Cambridge, 1961.

[Fal] K. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, 2nd edition, John Wiley & Sons, New York, 2003.

[Fef] C. Fefferman, *Pointwise convergence of Fourier series*, Ann. of Math. Ser. 2 **98** (1973), 551–571.

[Fe] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd edition, Vol. II, John Wiley & Sons, New York, 1971.

[Fi] D. Fishman, *Closed form continued fraction expansions of special quadratic irrationals*, Junior Thesis, Princeton University, 2003.

[Fol] G. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd edition, Pure and Applied Mathematics, Wiley-Interscience, New York, 1999.

[For] P. Forrester, *Log-gases and random matrices*, book in progress.

[Fou] E. Fouvry, *Sur la hauteur des points d'une certaine surface cubique singulière*. In *Nombre et répartition de points de hauteur bornée (Paris, 1996)*, Astérisque, (1999) no. 251, 31–49.

[FSV] P. J. Forrester, N. C. Snaith, and J. J. M. Verbaarschot, *Developments in Random Matrix Theory*. In *Random matrix theory*, J. Phys. A **36** (2003), no. 12, R1–R10.

[Fr] J. Franklin, *Mathematical Methods of Economics: Linear and Nonlinear Programming, Fixed-Point Theorem*, Springer-Verlag, New York, 1980.

[Ga] P. Garrett, *Making, Breaking Codes: An Introduction to Cryptography*, Prentice-Hall, Englewood Cliffs, NJ, 2000.

[Gau] M. Gaudin, *Sur la loi limite de l'espacement des valeurs propres d'une matrice aléatoire*, Nucl. Phys. **25** (1961) 447–458.

[Gel] A. O. Gelfond, *Transcendental and Algebraic Numbers*, Dover, New York, 1960.

[Gl] A. Gliga, *On continued fractions of the square root of prime numbers*, Junior Thesis, Princeton University, 2003.

[Gö] K. Gödel, *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*, Dover, New York, 1992.

[Gol1] D. Goldfeld, *The class number of quadratic fields and the conjectures of Birch and Swinnerton-Dyer*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 3, **4** (1976), 624–663.

[Gol2] D. Goldfeld, *The Elementary proof of the Prime Number Theorem, An Historical Perspective*. Pages 179–192 in *Number Theory, New York Seminar 2003*, eds. D. and G. Chudnovsky, M. Nathanson, Springer-Verlag, New York, 2004.

[Gold] L. Goldmakher, *On the limiting distribution of eigenvalues of large random regular graphs with weighted edges*, American Institute of Mathematics Summer REU, 2003.

[GV] D. A. Goldston and R. C. Vaughan, *On the Montgomery-Hooley asymptotic formula*. Pages 117–142 in *Sieve Methods, Exponential Sums and their Applications in Number Theory*, ed. G. R. H. Greaves, G. Harman, and M. N. Huxley, Cambridge University Press, Cambridge, 1996.

[GG] M. Golubitsky and V. Guillemin, *Stable Mappings and Their Singularities*, Graduate Texts in Mathematics, Vol. 14, Springer-Verlag, New York, 1973.

[Gou] X. Gourdon, *The 10^{13} first zeros of the Riemann zeta function, and zeros computation at very large height*, preprint. http://numbers.computation.free.fr/Constants/Miscellaneous/zetazeros1e13-1e24.pdf

[GKP] R. L. Graham, D. E. Knuth, and O. Patashnik, *Concrete Mathematics: A Foundation for Computer Science*, Addison-Wesley, Reading, MA, 1988.

[GK] A. Granville and P. Kurlberg, *Poisson statistics via the Chinese remainder theorem*, preprint.

[GT] A. Granville and T. Tucker, *It's as easy as abc*, Notices of the AMS **49** (2002), no. 10, 224–1231.

[GZ] B. Gross and D. Zagier, *Heegner points and derivatives of L-series*, Invent. Math. **84** (1986), no. 2, 225–320.

[Guy] R. Guy, *Unsolved Problems in Number Theory (Problem Books in Mathematics)*, 2nd edition, Springer-Verlag, New York, 1994.

[HM] C. Hammond and S. J. Miller, *Eigenvalue spacing distribution for the ensemble of real symmetric Toeplitz matrices*, Journal of Theoretical Probability **18** (2005), no. 3, 537–566.

[HL1] G. H. Hardy and J. E. Littlewood, *A new solution of Waring's problem*, Q. J. Math. **48** (1919), 272–293.

[HL2] G. H. Hardy and J. E. Littlewood, *Some problems of "Partitio Numerorum." A new solution of Waring's problem*, Göttingen Nach. (1920), 33–54.

[HL3] G. H. Hardy and J. E. Littlewood, *Some problems of "Partitio Numerorum." III. On the expression of a number as a sum of primes*, Acta Math. **44** (1923), 1–70.

[HL4] G. H. Hardy and J. E. Littlewood, *Some problems of "Partitio Numerorum." IV. Further researches in Waring's problem*, Math. Z. **23** (1925) 1–37.

[HR] G. H. Hardy and S. Ramanujan, *Asymptotic formulae in combinatorial analysis*, Proc. London Math. Soc. **17** (1918), 75–115.

[HW] G. H. Hardy and E. Wright, *An Introduction to the Theory of Numbers*, 5th edition, Oxford Science Publications, Clarendon Press, Oxford, 1995.

[Hata] R. Hata, *Improvement in the irrationality measures of π and π^2*, Proc. Japan. Acad. Ser. A Math. Sci. **68** (1992), 283–286.

[Ha1] B. Hayes, *Third Base: Three cheers for base 3!*, American Scientist **89** (2001), no. 6, 490–494.

[Ha2] B. Hayes, *The spectrum of Riemannium*, American Scientist **91** (2003), no. 4, 296–300.

[He] R. Heath-Brown, *The density of rational points on Cayley's cubic surface*, preprint.

[Hei] H. Heillbronn, *On the average length of a class of finite continued fractions*. In *Number Theory and Analysis (A collection of papers in honor of E. Landau)*, VEB Deutscher Verlag, Berlin, 1968.

[Hej] D. Hejhal, *On the triple correlation of zeros of the zeta function*, Internat. Math. Res. Notices (1994), no. 7, 294–302.

[Hil] D. Hilbert, *Beweis für die Darstellbarkeit der ganzen zahlen durch eine feste Anzahl n^{ter} Potenzen (Waringsches Problem)*, Mat. Annalen **67** (1909), 281–300.

[Hi1] T. Hill, *The first-digit phenomenon*, American Scientist **86** (1996), 358–363.

[Hi2] T. Hill, *A statistical derivation of the significant-digit law*, Statistical Science **10** (1996), 354–363.

[HS] M. Hindry and J. Silverman, *Diophantine Geometry: An Introduction*, Graduate Texts in Mathematics, Vol. 201, Springer-Verlag, New York, 2000.

[HSP] J. Hoffstein, J. H. Silverman and J. Pipher, *An Introduction to Mathematical Cryptography*.

[HJ] K. Hrbacek and T. Jech, *Introduction to Set Theory*, Pure and Applied Mathematics, Marcel Dekker, New York, 1984.

[Hua] Hua Loo Keng, *Introduction to Number Theory*, Springer-Verlag, New York, 1982.

[HuRu] C. Hughes and Z. Rudnick, *Mock Gaussian behaviour for linear statistics of classical compact groups*, J. Phys. A **36** (2003) 2919–2932.

[Hu] J. Hull, *Options, Futures, and Other Derivatives*, 5th edition, Prentice-Hall, Englewood Cliffs, NJ, 2002.

[IR] K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, Graduate Texts in Mathematics, Vol. 84, Springer-Verlag, New York, 1990.

[Iw] H. Iwaniec, *Topics in Classical Automorphic Forms*, Graduate Studies in Mathematics, Vol. 17, AMS, Providence, RI, 1997.

[IK] H. Iwaniec and E. Kowalski, *Analytic Number Theory*, AMS Colloquium Publications, Vol. 53, AMS, Providence, RI, 2004.

[ILS] H. Iwaniec, W. Luo, and P. Sarnak, *Low lying zeros of families of L-functions*, Inst. Hautes Études Sci. Publ. Math. **91** (2000), 55–131.

[IS1] H. Iwaniec and P. Sarnak, *Dirichlet L-functions at the central point*. Pages 941–952 in *Number Theory in Progress, (Zakopane-Kościelisko, 1997)*, Vol. 2, de Gruyter, Berlin, 1999.

[IS2] H. Iwaniec and P. Sarnak, *The non-vanishing of central values of automorphic L-functions and Landau-Siegel zeros*, Israel J. Math. **120** (2000), 155–177.

[JMRR] D. Jakobson, S. D. Miller, I. Rivin, and Z. Rudnick, *Eigenvalue spacings for regular graphs*. Pages 317–327 in *Emerging Applications of Number Theory (Minneapolis, 1996)*, The IMA Volumes in Mathematics and its Applications, Vol. 109, Springer, New York, 1999.

[J] N. Jacobson, *Basic Algebra I*, 2nd edition, W H Freeman & Co, San Francisco, 1985.

[Je] R. Jeffrey, *Formal Logic: Its Scope and Limits*, McGraw-Hill, New York, 1989.

[Ka] S. Kapnick, *Continued fraction of cubed roots of primes*, Junior Thesis, Princeton University, Fall 2002.

[Kar] A. Karlsson, *Applications of heat kernels on Abelian groups: $\zeta(2n)$, quadratic reciprocity, Bessel integral*, preprint.

[KS1] N. Katz and P. Sarnak, *Random Matrices, Frobenius Eigenvalues and Monodromy*, AMS Colloquium Publications, Vol. 45, AMS, Providence, RI, 1999.

[KS2] N. Katz and P. Sarnak, *Zeros of zeta functions and symmetries*, Bull. AMS **36** (1999), 1–26.

[KeSn] J. P. Keating and N. C. Snaith, *Random matrices and L-functions*. In *Random Matrix Theory*, J. Phys. A **36** (2003), no. 12, 2859–2881.

[Kei] J. B. Keiper, *Power series expansions of Riemann's ξ function*, Math. Comp. **58**(1992), 765-773.

[Kel] D. Kelley, *Introduction to Probability*, Macmillan Publishing Company, London, 1994.

[Kh] A. Y. Khinchin, *Continued Fractions*, 3rd edition, University of Chicago Press, Chicago, 1964.

[KSS] D. Kleinbock, N. Shah, and A. Starkov,*Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory*. Pages 813–930 in *Handbook of Dynamical Systems*, Vol. 1A, North-Holland, Amsterdam, 2002.

[Kn] A. Knapp, *Elliptic Curves*, Princeton University Press, Princeton, NJ, 1992.

[Knu] D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, 3rd edition, Addison-Wesley, MA, 1997.

[Kob1] N. Koblitz, *Why study equations over finite fields?*, Math. Mag. **55** (1982), no. 3, 144–149.

[Kob2] N. Koblitz, *Elliptic curve cryptosystems*, Math. Comp. **48** (1987), no. 177, 203–209.

[Kob3] N. Koblitz, *A survey of number theory and cryptography*. Pages 217-239 in *Number Theory*, Trends in Mathematics, Birkhäuser, Basel, 2000.

[Ko] V. Kolyvagin, *On the Mordell-Weil group and the Shafarevich-Tate group of modular elliptic curves*. Pages 429-436 in *Proceedings of the International Congress of Mathematicians (Kyoto, 1990)*, vols. I and II, Math. Soc. Japan, Tokyo, 1991.

[KonMi] A. Kontorovich and S. J. Miller, *Benford's law, values of L-functions and the $3x + 1$ problem*, Acta Arith. **120** (2005), 269–297.

[KonSi] A. Kontorovich and Ya. G. Sinai, *Structure theorem for* (d, g, h)*-maps*, Bull. Braz. Math. Soc. (N.S.) 33 (2002), no. 2, 213–224.

[Kor] A. Korselt, *Probléme chinois*, L'intermédiaire math. **6** (1899), 143–143.

[Kos] T. Koshy, *Fibonacci and Lucas Numbers with Applications*, Wiley-Interscience, New York, 2001

[Kua] F. Kuan, *Digit distribution in the continued fraction of* $\zeta(n)$, Junior Thesis, Princeton University, Fall 2002.

[KN] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, John Wiley & Sons, New York, 1974.

[KR] P. Kurlberg and Z. Rudnick, *The distribution of spacings between quadratic residues*, Duke Math. J. **100** (1999), no. 2, 211–242.

[Ku] R. Kuzmin, *Ob odnoi zadache Gaussa*, Doklady Akad. Nauk, Ser. A (1928), 375–380.

[Lag1] J. Lagarias, *The* $3x + 1$ *problem and its generalizations*. Pages 305-334 in *Organic mathematics (Burnaby, BC, 1995)*, CMS Conf. Proc., vol. 20, AMS, Providence, RI, 1997.

[Lag2] J. Lagarias, *The 3x+1 problem: An annotated bibliography*, preprint.

[LaSo] J. Lagarias and K. Soundararajan, *Benford's Law for the* $3x + 1$ *function*, preprint.

[La1] S. Lang, *Diophantine Geometry*, Interscience Publishers, New York, 1962.

[La2] S. Lang, *Introduction to Diophantine Approximations*, Addison-Wesley, Reading, MA, 1966.

[La3] S. Lang, *Undergraduate Algebra*, 2nd edition, Springer-Verlag, New York, 1986.

[La4] S. Lang, *Calculus of Several Variables*, Springer-Verlag, New York, 1987.

[La5] S. Lang, *Undergraduate Analysis*, 2nd edition, Springer-Verlag, New York, 1997.

[La6] S. Lang, *Complex Analysis*, Graduate Texts in Mathematics, Vol. 103, Springer-Verlag, New York, 1999.

[LT] S. Lang and H. Trotter, *Continued fractions for some algebraic numbers*, J. Reine Angew. Math. **255** (1972), 112–134.

[LF] R. Larson and B. Farber, *Elementary Statistics: Picturing the World*, Prentice-Hall, Englewood Cliffs, NJ, 2003.

[LP] R. Laubenbacher and D. Pengelley, *Gauss, Eisenstein, and the "third" proof of the quadratic reciprocity theorem: Ein kleines Schauspiel*, Math. Intelligencer 16 (1994), no. 2, 67–72.

[Law1] J. Law, *Kuzmin's theorem on algebraic numbers*, Junior Thesis, Princeton University, Fall 2002.

[Law2] J. Law, *The circle method on the binary Goldbach conjecture*, Junior Thesis, Princeton University, Spring 2003.

[Leh] R. Lehman, *First order spacings of random matrix eigenvalues*, Junior Thesis, Princeton University, Spring 2000.

[LS] H. Lenstra and G. Seroussi, *On hats and other covers*, 2002, preprint.

[Le] P. Lévy, *Sur les lois de probabilit$\acute{e}$ dont dependent les quotients complets et incomplets d'une fraction continue*, Bull. Soc. Math. **57** (1929), 178–194.

[XLi] X.-J. Li, *The positivity of a sequence of numbers and Riemann hypothesis*, J. Number Theory **65** (1997), 325-333.

[LU] C. Liaw and H. Úlfarsson, *Transcendence of e and π*, class notes for Math 252 (Graduate Algebra), Brown University, Spring 2006.

[Lidl] R. Lidl, *Mathematical aspects of cryptanalysis*. Pages 86–97 in *Number Theory and Cryptography (Sydney, 1989)*, London Mathematical Society Lecture Note Series, vol. 154, Cambridge University Press, Cambridge, 1990.

[Li] R. Lipshitz, *Numerical results concerning the distribution of $\{n^2\alpha\}$*, Junior Thesis, Princeton University, Spring 2000.

[Liu] Y. Liu, *Statistical behavior of the eigenvalues of random matrices*, Junior Thesis, Princeton University, Spring 2000.

[Mah] K. Mahler, *Arithmetische Eigenschaften einer Klasse von Dezimalbrüchen*, Amsterdam Proc. Konin. Neder. Akad. Wet. **40** (1937), 421–428.

[Ma] E. S. Mahmoodian, *Mathematical Olympiads in Iran*, Vol. I, Sharif University Press, Tehran, Iran, 2002.

[Man] B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1982.

[Mar] J. Marklof, *Almost modular functions and the distribution of n^2x modulo one*, Int. Math. Res. Not. (2003), no. 39, 2131–2151.

[MaMc] R. Martin and W. McMillen, *An elliptic curve over $\mathbb{Q}$ with rank at least* 24, Number Theory Listserver, May 2000.

[MMS] A. Massey, S. J. Miller, and J. Sinsheimer, *Eigenvalue spacing distribution for the ensemble of real symmetric palindromic Toeplitz matrices*, to appear in the Journal of Theoretical Probability.

[Maz1] B. Mazur, *Modular curves and the Eisenstein ideal*, IHES Publ. Math. **47** (1977), 33–186.

[Maz2] B. Mazur, *Rational isogenies of prime degree (with an appendix by D. Goldfeld)*, Invent. Math. **44** (1978), no. 2, 129–162.

[Maz3] B. Mazur, *Number Theory as Gadfly*, Amer. Math. Monthly, **98** (1991), 593–610.

[McK] B. McKay, *The expected eigenvalue distribution of a large regular graph*, Linear Algebra Appl. **40** (1981), 203–216.

[McW] B. McKay and N. Wormald, *The degree sequence of a random graph. I. The models,* Random Structures Algorithms **11** (1997), no. 2, 97–117.

[Meh1] M. Mehta, *On the statistical properties of level spacings in nuclear spectra*, Nucl. Phys. **18** (1960), 395–419.

[Meh2] M. Mehta, *Random Matrices*, 2nd edition, Academic Press, Boston, 1991.

[Met] N. Metropolis, *The beginning of the Monte Carlo method,* Los Alamos Science, No. 15, Special Issue (1987), 125–130.

[MU] N. Metropolis and S. Ulam, *The Monte Carlo method*, J. Amer. Statist. Assoc. **44** (1949), 335–341.

[Mic1] M. Michelini, *Independence of the digits of continued fractions*, Junior Thesis, Princeton University, Fall 2002.

[Mic2] M. Michelini, *Kuzmin's extraordinaty zero measure set*, Senior Thesis, Princeton University, Spring 2004.

[Mi1] N. Miller, *Various tendencies of non-Poissonian distributions along subsequences of certain transcendental numbers*, Junior Thesis, Princeton University, Fall 2002.

[Mi2] N. Miller, *Distribution of eigenvalue spacings for band-diagonal matrices*, Junior Thesis, Princeton University, Spring 2003.

[Mill] S. D. Miller, *A simpler way to show $\zeta(3)$ is irrational*, preprint.

[Mil1] S. J. Miller, 1- *and* 2-*level densities for families of elliptic curves: Evidence for the underlying group symmetries*, Compositio Mathematica **140** (2004), no. 4, 952–992.

[Mil2] S. J. Miller, *Density functions for families of Dirichlet characters*, preprint.

[Mil3] S. J. Miller, *The arithmetic mean and geometric inequality*, Class Notes from Math 187/487, The Ohio State University, Fall 2003.

[Mil4] S. J. Miller, *Differentiating identities*, Class Notes from Math 162: Statistics, Brown University, Spring 2005.

[Mil5] S. J. Miller, *The Pythagorean won-loss formula in baseball*, preprint.

[Mil6] S. J. Miller, *Investigations of zeros near the central point of elliptic curve L-functions*, to appear in Experimental Mathematics.

[Mil7] S. J. Miller, *Die battles and order statistics*, Class Notes from Math 162: Statistics, Brown University, Spring 2006.

[Mil8] S. J. Miller, *Beyond the Pigeon-Hole Principle: Many pigeons in the same box*, Class Notes from Math 162: Statistics, Brown University, Spring 2006.

[MN] S. J. Miller and M. Nigrini, *Order Statistics and Shifted Almost Benford Behavior*, preprint.

[M] V. Miller, *Use of elliptic curves in cryptography*. Pages 417–426 in *Advances in cryptology – CRYPTO '85 (Santa Barbara, CA, 1985)*, Lecture Notes in Computer Science, Vol. 218, Springer-Verlag, Berlin, 1986.

[Milne] J. S. Milne, *Elliptic Curves*, course notes.

[Min] S. Minteer, *Analysis of Benford's law applied to the $3x+1$ problem*, Number Theory Working Group, The Ohio State University, 2004.

[Mon1] H. Montgomery, *Primes in arithmetic progression*, Michigan Math. J. **17** (1970), 33–39.

[Mon2] H. Montgomery, *The pair correlation of zeros of the zeta function*. Pages 181–193 in *Analytic Number Theory*, Proceedings of Symposia in Pure Mathematics, vol. 24, AMS, Providence, RI, 1973.

[MoMc] D. Moore and G. McCabe, *Introduction to the Practice of Statistics*, W. H. Freeman and Co., London, 2003.

[MS] H. Montgomery and K. Soundararajan, *Beyond pair correlation*. Pages 507–514 in *Paul Erdös and His Mathematics, I (Budapest, 1999)*, Bolyai Society Mathematical Studies, Vol. 11, János Bolyai Math. Soc., Budapest, 2002.

[MW] C. J. Moreno and S. S. Wagstaff, Jr., *Sums of Squares of Integers*, Chapman and Hall, 2006.

[Moz1] C. J. Mozzochi, *An analytic sufficiency condition for Goldbach's conjecture with minimal redundancy*, Kyungpook Math. J. **20** (1980), no. 1, 1–9.

[Moz2] C. J. Mozzochi, *The Fermat Diary*, AMS, Providence, RI, 2000.

[Moz3] C. J. Mozzochi, *The Fermat Proof*, Trafford Publishing, Victoria, 2004.

[Mu1] R. Murty, *Primes in certain arithmetic progressions*, Journal of the Madras University, (1988), 161–169.

[Mu2] R. Murty, *Problems in Analytic Number Theory*, Springer-Verlag, New York, 2001.

[MM] M. R. Murty and V. K. Murty, *Non-Vanishing of L-Functions and Applications*, Progress in Mathematics, vol. 157, Birkhäuser, Basel, 1997.

[NS] K. Nagasaka and J. S. Shiue, *Benford's law for linear recurrence sequences*, Tsukuba J. Math. **11** (1987), 341–351.

[Nar] W. Narkiewicz, *The Development of Prime Number Theory*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2000.

[Na] M. Nathanson, *Additive Number Theory: The Classical Bases*, Graduate Texts in Mathematics, Springer-Verlag, New York, 1996.

[NT] J. von Neumann and B. Tuckerman, *Continued fraction expansion of* $2^{1/3}$, Math. Tables Aids Comput. **9** (1955), 23–24.

[Ni1] T. Nicely, *The pentium bug*, http://www.trnicely.net/pentbug/pentbug.html

[Ni2] T. Nicely, *Enumeration to* 10^{14} *of the Twin Primes and Brun's Constant*, Virginia J. Sci. **46** (1996), 195–204.

[Nig1] M. Nigrini, *Digital Analysis and the Reduction of Auditor Litigation Risk*. Pages 69–81 in *Proceedings of the 1996 Deloitte & Touche / University of Kansas Symposium on Auditing Problems*, ed. M. Ettredge, University of Kansas, Lawrence, KS, 1996.

[Nig2] M. Nigrini, *The Use of Benford's Law as an Aid in Analytical Procedures*, Auditing: A Journal of Practice & Theory, **16** (1997), no. 2, 52–67.

[NZM] I. Niven, H. Zuckerman, and H. Montgomery, *An Introduction to the Theory of Numbers*, 5th edition, John Wiley & Sons, New York, 1991.

[Nov] T. Novikoff, *Asymptotic behavior of the random* 3*-regular bipartite graph*, Undergraduate Mathematics Laboratory report, Courant Institute, NYU, 2002.

[Ny] J. E. Nymann, *An Application of Diophantine Approximation*, The American Mathematical Monthly **76** (1969), no. 6, 668–671.

[Od1] A. Odlyzko, *On the distribution of spacings between zeros of the zeta function*, Math. Comp. **48** (1987), no. 177, 273–308.

[Od2] A. Odlyzko, *The* 10^{22}*-nd zero of the Riemann zeta function*. Pages 139–144 in *Proceedings of the Conference on Dynamical, Spectral and Arithmetic Zeta Functions*, ed. M. van Frankenhuysen and M. L. Lapidus, Contemporary Mathematics Series, AMS, Providence, RI, 2001.

[Ok] T. Okano, *A note on the transcendental continued fractions*, Tokyo J. Math **10** (1987), no. 1, 151–156.

[Ol] T. Oliveira e Silva, *Verification of the Goldbach conjecture up to* $6 \cdot 10^{16}$, NMBRTHRY@listserv.nodak.edu mailing list, Oct. 3, 2003, http://listserv.nodak.edu/scripts/wa.exe?A2=ind0310&L=nmbrthry&P=168 and http://www.ieeta.pt/$\sim$tos/goldbach.html

[Ols] L. Olsen, *Extremely non-normal continued fractions*, Acta Arith. **108** (2003), no. 2, 191–202.

[Pi] R. G. E. Pinch, *The Carmichael numbers up to* 10^{18}, preprint, http://arxiv.org/abs/math.NT/0604376.

[Pol] G. Polya, *Heuristic reasoning in the theory of numbers*, Amer. Math. Monthly **66** (1959) 375–384.

[vdP1] A. van der Poorten, *An introduction to continued fractions*. Pages 99-138 in *Diophantine Analysis (Kensington, 1985)*, London Mathematical Society Lecture Note Series, Vol. 109, Cambridge University Press, Cambridge, 1986.

[vdP2] A. van der Poorten, *Notes on continued fractions and recurrence sequences*. Pages 86–97 in *Number theory and cryptography (Sydney, 1989)*, London Mathematical Society Lecture Note Series, Vol. 154, Cambridge University Press, Cambridge, 1990.

[vdP3] A. van der Poorten, *Continued fractions of formal power series*. Pages 453–466 in *Advances in Number Theory (Kingston, ON, 1991)*, Oxford Science Publications, Oxford University Press, New York, 1993.

[vdP4] A. van der Poorten, *Fractions of the period of the continued fraction expansion of quadratic integers*, Bull. Austral. Math. Soc. **44** (1991), no. 1, 155–169.

[vdP5] A. van der Poorten, *Continued fraction expansions of values of the exponential function and related fun with continued fractions*, Nieuw Arch. Wisk. (4) **14** (1996), no. 2, 221–230.

[vdP6] A. van der Poorten, *Notes on Fermat's Last Theorem*, Canadian Mathematical Society Series of Monographs and Advanced Texts, Wiley-Interscience, New York, 1996.

[PS1] A. van der Poorten and J. Shallit, *Folded continued fractions*, J. Number Theory **40** (1992), no. 2, 237–250.

[PS2] A. van der Poorten and J. Shallit, *A specialised continued fraction*, Canad. J. Math. **45** (1993), no. 5, 1067–1079.

[Po] C. Porter (editor), *Statistical Theories of Spectra: Fluctuations*, Academic Press, New York, 1965.

[Py] R. Pyke, *Spacings*, J. Roy. Statist. Soc. Ser. B **27** (1965), 395–449.

[QS1] R. Qian and D. Steinhauer, *Rational relation conjectures*, Junior Thesis, Princeton University, Fall 2003.

[QS2] R. Qian and D. Steinhauer, *Eigenvalues of weighted random graphs*, Junior Thesis, Princeton University, Spring 2003.

[Rai] R. A. Raimi, *The first digit problem*, Amer. Math. Monthly **83** (1976), no. 7, 521–538.

[Ra] K. Ramachandra, *Lectures on Transcendental Numbers*, Ramanujan Institute, Madras, 1969.

[Re] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw-Hill, New York, 1965.

[Ric] P. Richter, *An investigation of expanders and ramanujan graphs along random walks of cubic bipartite graphs*, Junior Thesis, Princeton University, Spring 2001.

[RDM] R. D. Richtmyer, M. Devaney, and N. Metropolis, *Continued fraction of algebraic numbers*, Numer. Math. **4** (1962), 68–84.

[Rie] H. J. J. te Riele, *On the sign of the difference* $\pi(x) - \text{Li}(x)$, Mathematics of Computation **48** (1987), no. 177, 323–328.

[Ri] G. F. B. Riemann, *Über die Anzahl der Primzahlen unter einer gegebenen Grösse*, Monatsber. Königl. Preuss. Akad. Wiss. Berlin, Nov. 1859, 671–680 (see [Ed] for an English translation).

[RSA] R. Rivest, A. Shamir, and L. Adleman, *A method for obtaining digital signatures and public key cryptosystems*, Comm. ACM **21** (1978), 120–126.

[Roc] D. Rockmore, *Stalking the Riemann Hypothesis: The Quest to Find the Hidden Law of Prime Numbers*, Pantheon, New York, 2005.

[Ro] K. Roth, *Rational approximations to algebraic numbers*, Mathematika **2** (1955), 1–20.

[Rub1] M. Rubinstein, *A simple heuristic proof of Hardy and Littlewood's conjecture B*, Amer. Math. Monthly **100** (1993), no. 5, 456–460.

[Rub2] M. Rubinstein, *Low-lying zeros of* L*-functions and random matrix theory*, Duke Math. J. **109** (2001), no. 1, 147–181.

[RubSa] M. Rubinstein and P. Sarnak, *Chebyshev's bias*, Experiment. Math. **3** (1994), no. 3, 173–197.

[Rud] W. Rudin, *Principles of Mathematical Analysis*, 3rd edition, International Series in Pure and Applied Mathematics, McGraw-Hill, New York, 1976.

[RS] Z. Rudnick and P. Sarnak, *Zeros of principal* L*-functions and random matrix theory*, Duke J. of Math. **81** (1996), 269–322.

[RS2] Z. Rudnick and P. Sarnak, *The pair correlation function of fractional parts of polynomials*, Comm. Math. Phys. **194** (1998), no. 1, 61–70.

[RSZ] Z. Rudnick, P. Sarnak, and A. Zaharescu, *The distribution of spacings between the fractional parts of* $n^2\alpha$, Invent. Math. **145** (2001), no. 1, 37–57.

[RZ1] Z. Rudnick and A. Zaharescu, *A metric result on the pair correlation of fractional parts of sequences*, Acta Arith. **89** (1999), no. 3, 283–293.

[RZ2] Z. Rudnick and A. Zaharescu, *The distribution of spacings between fractional parts of lacunary sequences*, Forum Math. **14** (2002), no. 5, 691–712.

[Sai] F. Saidak, *A new proof of Euclid's theorem*, Amer. Math. Monthly **113** (2006), no. 10, 937–938.

[Sar] P. Sarnak *Some applications of modular forms*, Cambridge Trusts in Mathemetics, Vol. 99, Cambridge University Press, Cambridge, 1990.

[Sch] D. Schmidt, *Prime Spacing and the Hardy-Littlewood Conjecture B*, Junior Thesis, Princeton University, Spring 2001.

[Sc] P. Schumer, *Mathematical Journeys*, Wiley-Interscience, John Wiley & Sons, New York, 2004.

[Se] J. P. Serre, *A Course in Arithmetic*, Springer-Verlag, New York, 1996.

[Sh] A. Shidlovskii, *Transcendental Numbers*, Walter de Gruyter & Co., New York, 1989.

[ShTa] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*, AMS, Providence, RI, 1943.

[Sil1] J. Silverman, *The Arithmetic of Elliptic Curves*, Graduate Texts in Mathematics, Vol. 106, Springer-Verlag, New York, 1986.

[Sil2] J. Silverman, *A Friendly Introduction to Number Theory*, 2nd edition, Prentice-Hall, Englewood Cliffs, NJ, 2001.

[ST] J. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer-Verlag, New York, 1992.

[Si] B. Simon, *The classical moment problem as a self-adjoint finite difference operator*, Adv. Math. **137** (1998), no. 1, 82–203.

[SM] S. Simon and A. Moustakas, *Eigenvalue density of correlated complex random Wishart matrices*, Bell Labs Technical Memo, 2004.

[Sk] S. Skewes, *On the difference* $\pi(x) - \text{Li}(x)$, J. London Math. Soc. **8** (1933), 277–283.

[Sl] N. Sloane, *On-Line Encyclopedia of Integer Sequences*, http://www.research.att.com/∼njas/sequences/Seis.html

[Sn] N. Snaith, *Derivatives of random matrix characteristic polynomials with applications to elliptic curves*, J. Phys. A **38** (2005), no. 48, 10345–10360.

[So] K. Soundararajan, *Small gaps between prime numbers: The work of Goldston-Pintz-Yildirim*, Bull. of the AMS **44** (2007), no. 1, 1–18.

[SS1] E. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press, Princeton, NJ, 2003.

[SS2] E. Stein and R. Shakarchi, *Complex Analysis*, Princeton University Press, Princeton, NJ, 2003.

[SS3] E. Stein and R. Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*, Princeton University Press, Princeton, NJ, 2005.

[StTa] I. Stewart and D. Tall, *Algebraic Number Theory*, 2nd edition, Chapman & Hall, London, 1987.

[St] Strang, *Linear Algebra and Its Applications*, 3rd edition, Wellesley-Cambridge Press, Wellesley, MA 1998.

[Str] K. Stromberg, *The Banach-Tarski paradox*, Amer. Math. Monthly **86** (1979), no. 3, 151–161.

[Sz] P. Szüsz, *On the length of continued fractions representing a rational number with given denominator*, Acta Arithmetica **37** (1980), 55–59.

[Ta] C. Taylor, *The Gamma function and Kuzmin's theorem*, Junior Thesis, Princeton University, Fall 2002.

[TW] R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. Math. **141** (1995), 553–572.

[TrWi] C. Tracy and H. Widom, *Correlation functions, cluster functions, and spacing distributions for random matrices*, J. Statist. Phys. **92** (1998), no. 5–6, 809–835.

[Te] G. Tenenbaum, *Introduction to Analytic and Probabilistic Number Theory*, Cambridge University Press, Cambridge, 1995.

[Ti] E. C. Titchmarsh, *The Theory of the Riemann Zeta-function*, revised by D. R. Heath-Brown, Oxford University Press, Oxford, 1986.

[Va] R. C. Vaughan, *On a variance associated with the distribution of primes in arithmetic progression*, Proc. London Math. Soc. (3) **82** (2001), 533–553.

[VW] R. C. Vaughan and T. D. Wooley, *Waring's problem: a survey*. Pages 301–340 in *Number Theory for the Millennium, III (Urbana, IL, 2000)*, A. K. Peters, Natick, MA, 2002.

[Vin1] I. Vinogradov, *Representation of an odd number as the sum of three primes*, Doklady Akad. Nauk SSSR **15** (1937), no. 6–7, 291–294.

[Vin2] I. Vinogradov, *Some theorems concerning the theory of primes*, Mat. Sbornik **2** (1937), no. 44, 179–195.

[Vo] A. Voros, *A sharpening of Li's criterion for the Riemann hypothesis*, preprint.

[VG] W. Voxman and R. Goetschel, Jr., *Advanced Calculus*, Mercer Dekker, New York, 1981.

[Wa] L. Washington, *Elliptic Curves: Number Theory and Cryptography*, Chapman & Hall / CRC, New York, 2003.

[Wed] S. Wedeniwski, *ZetaGrid*, http://www.zetagrid.net

[Wei1] A. Weil, *Numbers of Solutions of Equations in Finite Fields*, Bull. Amer. Math. Soc. **14** (1949), 497–508.

[Wei2] A. Weil, *Prehistory of the zeta-function*. Pages 1–9 in *Number Theory, Trace Formulas and Discrete Groups (Oslo, 1987)*, Academic Press, Boston, 1989.

[Weir] B. Weir, *The local behavior of Germain primes*, Undergraduate Mathematics Laboratory report, Courant Institute, NYU, 2002.

[We] E. Weisstein, *MathWorld — A Wolfram Web Resource*, http://mathworld.wolfram.com

[Weyl] H. Weyl, *The Classical Groups: Their Invariants and Representations*, Princeton University Press, Princeton, NJ, 1946.

[Wh] E. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies: With an Introduction to the Problem of Three Bodies*, Dover, New York, 1944.

[WW] E. Whittaker and G. Watson, *A Course of Modern Analysis*, 4th edition, Cambridge University Press, Cambridge, 1996.

[Wig1] E. Wigner, *On the statistical distribution of the widths and spacings of nuclear resonance levels*, Proc. Cambridge Philo. Soc. **47** (1951), 790–798.

[Wig2] E. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions*, Ann. of Math. **2** (1955), no. 62, 548–564.

[Wig3] E. Wigner, *Statistical Properties of real symmetric matrices*. Pages 174–184 in *Canadian Mathematical Congress Proceedings*, University of Toronto Press, Toronto, 1957.

[Wig4] E. Wigner, *Characteristic vectors of bordered matrices with infinite dimensions. II*, Ann. of Math. Ser. 2 **65** (1957), 203–207.

[Wig5] E. Wigner, *On the distribution of the roots of certain symmetric matrices*, Ann. of Math. Ser. 2 **67** (1958), 325–327.

[Wi] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. Math. **141** (1995), 443–551.

[Wilf] H. Wilf, *Algorithms and Complexity*, 2nd edition, A. K. Peters, Natick, MA, 2002.

[Wir] E. Wirsing, *On the theorem of Gauss-Kuzmin-Lévy and a Frobenius-type theorem for function spaces*, Acta Arith. **24** (1974) 507–528.

[Wis] J. Wishart, *The generalized product moment distribution in samples from a normal multivariate population*, Biometrika **20 A** (1928), 32–52.

[Wor] N. C. Wormald, *Models of random regular graphs*. Pages 239–298 in *Surveys in combinatorics, 1999 (Canterbury)* London Mathematical Society Lecture Note Series, vol. 267, Cambridge University Press, Cambridge, 1999.

[Wo] T. Wooley, *Large improvements in Waring's problem*, Ann. of Math. (2), **135** (1992), no. 1, 131–164.

[Za] I. Zakharevich, *A generalization of Wigner's law*, preprint.

[Zu] W. Zudilin, *One of the numbers* $\zeta(5), \zeta(7), \zeta(9), \zeta(11)$ *is irrational*, Uspekhi Mat. Nauk **56** (2001), 149-150.

[Zy] A. Zygmund, *Trigonometrical Series*, vols. I and II, Cambridge University Press, Cambridge, 1968.

Index