Chapter Eight

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 §8.4 and proved in §11.5.

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 9 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 9 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 7) 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 10.

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 $\beta_1,\dots,\beta_N$:

$$0 \le \beta_1 \le \beta_2 \le \cdots \le \beta_N. \tag{8.1}$$

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 12, and then in Part 5 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!

8.1 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 §8.2.3. We assume the reader is familiar with elementary set operations and countable sets (see §5.2).

8.1.1 Introduction

Definition 8.1.1 (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 8.1.2. *Flip a coin 3 times. The possible outcomes are*

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}. \tag{8.2}$$

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 8.1.3 (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 8.1.4 (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 8.1.2, 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

$$X_i(\omega) \;=\; \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ toss is a head} \\ 0 & \text{if the } i^{\text{th}} \text{ toss is a tail.} \end{cases} \tag{8.3}$$

Note that

$$X(\omega) \;=\; X_1(\omega) + X_2(\omega) + X_3(\omega). \tag{8.4}$$

Remark 8.1.5 (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 $x \in \{1, \ldots, 6\}$. 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 8.1.6 (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.

8.1.2 Events

Definition 8.1.7 (Events)**.** *We call a subset $A \subset \Omega$ an event, and we write*

$$\text{Prob}(A) \;=\; \sum_{\omega \in A} p(\omega). \tag{8.5}$$

Note each outcome is also an event.

Definition 8.1.8 (Range of X)**.** *The range of a random variable X is the set of values it attains, denoted $X(\Omega)$:*

$$X(\Omega) \;=\; \{r \in \mathbb{R} : \exists \omega \in \Omega \text{ with } X(\omega) = r\}. \tag{8.6}$$

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 8.1.9. *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:

$$A_i \;=\; \{\omega \in \Omega : X(\omega) = x_i\} \;\subset\; \Omega. \tag{8.7}$$

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 8.1.10 (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

$$\widetilde{p}(x_i) \;=\; \sum_{\substack{\omega \in \Omega \\ X(\omega) = x_i}} p(\omega). \tag{8.8}$$

Remark 8.1.11 (Important)**.** In a convenient abuse of notation, we often write

$$p(x_i) \;=\; p(X(\omega) = x_i) \;=\; \text{Prob}(\omega \in \Omega : X(\omega) = x_i). \tag{8.9}$$

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 8.1.14 illustrate.

Example 8.1.12. *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 8.1.13. *Calculate the probabilities of the events $C_2, \ldots, C_{12}$ for Example 8.1.12.*

Example 8.1.14 (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:*

$$1_A(\omega) \;=\; \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases} \tag{8.10}$$

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

Definition 8.1.15 (Complements)**.** *The complement of a set $A \subset \Omega$ is the set of all $\omega \notin A$. We denote this by A^c:*

$$A^c \;=\; \{\omega : \omega \in \Omega, \omega \notin A\}. \tag{8.11}$$

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

$$X_A(\omega) \;=\; \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c. \end{cases} \tag{8.12}$$

Lemma 8.1.16. *Consider an outcome space Ω with outcomes ω and probability function p. Let $A \subset \Omega$ be an event. Then*

$$p(A) \;=\; 1 - p(A^c). \tag{8.13}$$

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

Exercise 8.1.17. *Prove Lemma 8.1.16. Consider 100 tosses of a fair coin. What is the probability that at least three tosses are heads?*

Exercise$^{(hr)}$ 8.1.18. *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 8.1.19. *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 8.1.20 (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

$$x \;=\; p + rp + r^2p + \cdots \;=\; p\sum_{n=0}^{\infty} r^n. \tag{8.14}$$

However, we also know that

$$x \;=\; p + (1-p)(1-q)x \;=\; p + rx. \tag{8.15}$$

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 (8.14) *becomes*

$$\sum_{n=0}^{\infty} r^n \;=\; \frac{1}{1-r}, \tag{8.16}$$

the geometric series formula!

Exercise$^{(h)}$ 8.1.21. *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$^{(h)}$ 8.1.22 (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 8.1.23. Exercises 8.1.20 and 8.1.22 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(hr) 8.1.24. *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.*

8.1.3 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

$$\Omega \ = \ \{\omega = (\omega_1, \omega_2) : \omega_1 \in \Omega_1 \text{ and } \omega_2 \in \Omega_2\}, \tag{8.17}$$

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

$$p(\omega) \ = \ p_1(\omega_1) \cdot p_2(\omega_2) \text{ if } \omega = (\omega_1, \omega_2). \tag{8.18}$$

Exercise 8.1.25. *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 8.1.26. *What is* $\text{Prob}(X(\omega) = 2)$*? There are* 6 *pairs with first roll* 2*:* $(2,1), (2,2), \ldots, (2,6)$*. Each pair has probability* $\frac{1}{36}$*. Thus,* $\text{Prob}(X(\omega) = 2) = \frac{6}{36} = \frac{1}{6}$.

More generally we have

$$\text{Prob}\left(X(\omega) = x_i\right) \ = \ \sum_{\substack{\omega = (\omega_1, \omega_2) \\ X(\omega) = x_i}} p\left(\omega\right). \tag{8.19}$$

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

$$\text{Prob}(A) \ = \ \frac{\sum_{\omega \in A} p(w)}{\sum_{\omega \in \Omega} p(\omega)}. \tag{8.20}$$

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

$$\text{Prob}(A|B) \;=\; \frac{\sum_{\substack{\omega\in A\\ \omega\in B}} p(w)}{\sum_{\omega\in B} p(\omega)}. \tag{8.21}$$

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

Lemma 8.1.27. *If* $\text{Prob}(B) \neq 0$,

$$\text{Prob}(A|B) \;=\; \frac{\text{Prob}(A\cap B)}{\text{Prob}(B)}. \tag{8.22}$$

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

$$\begin{aligned} A &= \{(2,1),(2,2),(2,3),(2,4),(2,5),(2,6)\}\\ B &= \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}\\ A\cap B &= \{(2,5)\}\\ \text{Prob}(A|B) \;=\; \frac{\text{Prob}(A\cap B)}{\text{Prob}(B)} \;&=\; \frac{\frac{1}{36}}{6\cdot\frac{1}{36}} \;=\; \frac{1}{6}. \end{aligned} \tag{8.23}$$

Exercise 8.1.28. *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*

$$p(\omega) \;=\; \begin{cases} \frac{1.5}{36} & \text{if } \omega_1 \text{ is even}\\ \frac{.5}{36} & \text{if } \omega_1 \text{ is odd.} \end{cases} \tag{8.24}$$

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 8.1.29. *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.*

8.1.4 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:

$$\text{Prob}(A|B) \;=\; \frac{\text{Prob}(A \cap B)}{\text{Prob}(B)} \;=\; \text{Prob}(A). \tag{8.25}$$

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

$$\text{Prob}(A|B) \;=\; \frac{1}{6} \;=\; \text{Prob}(A); \tag{8.26}$$

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

$$\text{Prob}(A|B) \;=\; 0 \;\neq\; \text{Prob}(A); \tag{8.27}$$

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,

$$\text{Prob}(A|B) \;=\; \frac{\text{Prob}(A \cap B)}{\text{Prob}(B)} \;=\; \text{Prob}(A), \tag{8.28}$$

we have

$$\text{Prob}(A \cap B) \;=\; \text{Prob}(A)\text{Prob}(B). \tag{8.29}$$

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

$$\text{Prob}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \;=\; \text{Prob}(A_{i_1})\text{Prob}(A_{i_2}) \cdots \text{Prob}(A_{i_l}). \tag{8.30}$$

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

Exercise 8.1.30. *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 8.1.31. *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 8.1.32 (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 8.1.33. *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 $\omega = (\omega_1, \ldots, \omega_N) \in \Omega$, 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$^{(hr)}$ 8.1.34 (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 8.1.35. *Redo the previous problem assuming that there are one-fourth as many people born on February 29^{th} as on any other day.*

Exercise$^{(hr)}$ 8.1.36. *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

$$\text{Prob(Player one wins)} \;=\; \frac{1}{k^{m+n}} \sum_{a=2}^{k} [a^m - (a-1)^m] \cdot (a-1)^n, \qquad (8.31)$$

which by the integral version of partial summation equals

$$\frac{1}{k^{m+n}} \left[k^m \cdot (k-1)^n - \int_1^k [u]^m \cdot n(u-1)^{n-1} du \right]. \qquad (8.32)$$

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

$$\text{Prob(Player one wins)} \;=\; \frac{m}{m+n} - \frac{m}{2(m+n-1)} \frac{n}{k}; \qquad (8.33)$$

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

8.1.5 Expectation

Definition 8.1.37 (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*

$$\overline{X} \;=\; \sum_i X(\omega_i) p(\omega_i). \qquad (8.34)$$

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

Exercise 8.1.38. *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 8.1.39. 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 8.1.40. *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*

$$\Omega \;=\; \{H, TH, TTH, TTTH, \dots\}. \qquad (8.35)$$

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 8.1.41 (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*

$$\mathbb{E}[X^k] \;=\; \sum_{\omega_i \in \Omega} X(\omega_i)^k \cdot p(\omega_i). \qquad (8.36)$$

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

Definition 8.1.42 (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 8.1.6). We define the k^{th} moment of p by*

$$p_k \;=\; \mathbb{E}[X^k] \;=\; \sum_{x\in\Omega} x^k p(x). \tag{8.37}$$

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 5; see §15.3.2 for more details.

Exercise 8.1.43. *Prove the zeroth moment of any probability distribution is 1.*

Lemma 8.1.44 (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*

$$P(x_i, y_j) \;:=\; \text{Prob}(X(\omega) = x_i, Y(\omega) = y_j). \tag{8.38}$$

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

Proof. First note

$$\text{Prob}(X(\omega) = x_i) \;=\; \sum_j \text{Prob}(X(\omega) = x_i, Y(\omega) = y_j) \;=\; \sum_j P(x_i, y_j). \tag{8.39}$$

Thus the expected value of the random variable X is

$$\mathbb{E}[X] \;=\; \sum_i x_i \sum_j P(x_i, y_j), \tag{8.40}$$

and similarly for the random variable Y. Therefore

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{(i,j)} (x_i + y_j) P(x_i, y_j) \\
&= \sum_i \sum_j x_i P(x_i, y_j) + \sum_i \sum_j y_j P(x_i, y_j) \\
&= \sum_i x_i \sum_j P(x_i, y_j) + \sum_j y_j \sum_i P(x_i, y_j) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned} \tag{8.41}$$

□

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 11.4.12.

Lemma 8.1.45 (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*

$$\mathbb{E}[a_1X_1 + \cdots + a_NX_N] = a_1\mathbb{E}[X_1] + \cdots + a_N\mathbb{E}[X_N]. \tag{8.42}$$

See §10.5.2 for an application of the linearity of expected values to investigating digits of continued fractions.

Exercise 8.1.46. *Prove Lemma 8.1.45.*

Lemma 8.1.47. *Let X and Y be independent random variables. Then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.*

Proof. From Definition 8.1.32, 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

$$\text{Prob}(A_i \cap B_j) \;=\; \text{Prob}(A_i)\text{Prob}(B_j) \;=\; p(x_i)q(y_j). \tag{8.43}$$

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,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_i \sum_j x_i y_j r(x_i, y_j) \\
&= \sum_i \sum_j x_i y_j p(x_i) q(y_j) \\
&= \sum_i x_i p(x_i) \cdot \sum_j y_j q(y_j) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y].
\end{aligned} \tag{8.44}$$

□

Exercise 8.1.48. *Find two random variables such that $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$.*

Exercise 8.1.49 (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$^{(hr)}$ 8.1.50. *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.*

8.1.6 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 8.1.51 (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*

$$\sigma_X^2 \;=\; \sum_i \left(X(\omega_i) - \mathbb{E}[X]\right)^2 p(\omega_i) \;=\; \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]. \tag{8.45}$$

Exercise 8.1.52. *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 (8.45) one often uses

Lemma 8.1.53. *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

$$\begin{aligned}
\sigma_X^2 &= \sum_i \left(X_i(\omega) - \mathbb{E}[X]\right)^2 p(\omega_i) \\
&= \sum_i (X_i(\omega)^2 - 2X_i(\omega)\mathbb{E}[X] + \mathbb{E}[X]^2) p(\omega_i) \\
&= \sum_i X_i(\omega)^2 p(\omega_i) - 2\mathbb{E}[X] \sum_i X_i(\omega) p(\omega_i) + \mathbb{E}[X]^2 \sum_i p(\omega_i) \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned} \tag{8.46}$$

□

The main result on variances is

Lemma 8.1.54 (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 8.1.45).

$$\begin{aligned}
\sigma_{X+Y}^2 &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[(X+Y)]^2 \\
&= \mathbb{E}[X^2 + 2XY + Y^2] - \left(\mathbb{E}[X] + \mathbb{E}[Y]\right)^2 \\
&= \left(\mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2]\right) - \left(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\right) \\
&= \left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right) + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\
&= \sigma_X^2 + \sigma_Y^2 + 2\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right).
\end{aligned} \tag{8.47}$$

By Lemma 8.1.47, 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 8.1.45 that the mean of the random variable $X_1 + \cdots + X_N$ is $N\mathbb{E}[X]$. What is the variance?

Lemma 8.1.55. *Notation as above,*

$$\sigma_{X_1+\cdots+X_N} = \sqrt{N}\sigma_X. \tag{8.48}$$

Exercise 8.1.56. *Prove Lemma 8.1.55.*

Lemma 8.1.57. *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*

$$\sigma^2_{aX+b} = a^2\sigma^2_X. \tag{8.49}$$

Exercise 8.1.58. *Prove 8.1.57.*

Note that if the random variable X has units of meters then the variance σ^2_X 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

$$\sum_i (x_i - \overline{X})p(x_i). \tag{8.50}$$

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

Exercise 8.1.59. *Show $\sum_i (x_i - \overline{X})p(x_i) = 0$.*

This leads us to consider

$$\sum_i |x_i - \overline{X}|p(x_i). \tag{8.51}$$

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 8.1.60 (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)*

$$\sum_{i=1}^{n} (y_i - (at_i + b))^2 = \sum_{i=1}^{n} (\text{Observed}_i - \text{Predicted}_i)^2. \tag{8.52}$$

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 $a_1, \ldots, a_N$ 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 8.1.61. *Consider the best fit line from the Method of Least Squares (Exercise 8.1.60). 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 8.1.62 (Chebyshev's Inequality)**.** *Let X be a random variable with mean μ and finite variance σ^2. Prove Chebyshev's inequality:*

$$\text{Prob}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \tag{8.53}$$

where $\text{Prob}(|X - \mu| \geq a)$ 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 8.1.63. *Use Chebyshev's Theorem to bound the probability of tossing a fair coin* 10000 *times and observing at least* 6000 *heads.*

Exercise 8.1.64. *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 8.4.1) and obtain better estimates than those from Chebyshev's Theorem. See Exercise 8.4.3.

8.2 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 8.1.42. For example, for a discrete probability distribution p the mean is $\sum_i x_i p(x_i)$.

8.2.1 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 §8.2.4 we describe a Bernoulli trial to experimentally determine π!

Exercise 8.2.1. *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 8.1.14).*

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 8.1.45 and 8.1.55 allow one to calculate both quantities immediately, once one knows the mean and variance for a single occurrence (see Exercise 8.2.1).

Lemma 8.2.2. *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 8.2.2 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 §8.4.

Exercise 8.2.3. *Prove Lemma 8.2.2. 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

$$p(k) \;=\; \begin{cases} (1-p)^{k-1}\cdot p & \text{if } k\in\{1,2,\dots\} \\ 0 & \text{otherwise.} \end{cases} \tag{8.54}$$

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

$$\overline{Z} = \sum_{k=1}^{\infty} k(1-p)^{k-1}p \;=\; p\sum_{k=1}^{\infty} kq^{k-1}, \quad 0<q=1-p<1. \tag{8.55}$$

Consider the geometric series

$$f(q) \;=\; \sum_{k=0}^{\infty} q^k \;=\; \frac{1}{1-q}. \tag{8.56}$$

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

$$f'(q) \;=\; \sum_{k=0}^{\infty} kq^{k-1} \;=\; \frac{1}{(1-q)^2}. \tag{8.57}$$

Recalling $q=1-p$ and substituting yields

$$\overline{Z} = p\sum_{k=1}^{\infty} kq^{k-1} \;=\; \frac{p}{\left(1-(1-p)\right)^2} \;=\; \frac{1}{p}. \tag{8.58}$$

Remark 8.2.4. 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 8.2.5. *Calculate σ_Z^2.* Hint: *Differentiate $f(q)$ twice.*

Exercise 8.2.6. *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*

$$\sigma \;=\; \int_{-\infty}^{\infty} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}}\,dx. \tag{8.59}$$

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 $7!!=7\cdot5\cdot3\cdot1$ and $6!!=6\cdot4\cdot2$).

Exercise 8.2.7. *The even moments of the Gaussian (see Exercise 8.2.6) 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 §16.2.2, where we study the eigenvalues of real symmetric Toeplitz matrices.*

8.2.2 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 8.2.2 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

$$\begin{aligned} p_N(k) &= \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N!}{k!(N-k)!} \frac{\lambda^k}{N^k} \left(1 - \frac{\lambda}{N}\right)^{N-k} \\ &= \frac{N \cdot (N-1) \cdots (N-k+1)}{N \cdot N \cdots N} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^{N} \left(1 - \frac{\lambda}{N}\right)^{-k} \\ &= 1 \cdot \left(1 - \frac{1}{N}\right) \cdots \left(1 - \frac{k-1}{N}\right) \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^{N} \left(1 - \frac{\lambda}{N}\right)^{-k}. \end{aligned} \tag{8.60}$$

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 §5.4 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 12.

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

Exercise$^{(h)}$ 8.2.9. *Calculate the mean and variance for the Poisson Distribution.*

8.2.3 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

$$p_{N,S}(x) = \left(1 - \frac{\lambda}{N}\right)^{\lceil Nx \rceil - 1} \cdot \left(\frac{\lambda}{N}\right)^{1}. \tag{8.61}$$

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

Exercise 8.2.10. *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 8.2.11 (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

$$p_S(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0; \end{cases} \tag{8.62}$$

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 12 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 8.2.12 (**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 $[c, d] \subset [a, b]$,*

$$\text{Prob}([c, d]) = \int_c^d p(x)dx = \frac{d - c}{b - a}. \tag{8.63}$$

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 8.2.13 (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 8.2.14 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 §8.4 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

$$I \;=\; \int_{-\infty}^{\infty} e^{-x^2}\,dx. \tag{8.64}$$

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

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-x^2}dx \cdot \int_{-\infty}^{\infty} e^{-y^2}dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2-y^2}dxdy \\ &= \int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2} r\,dr \\ &= 2\pi \cdot \left[-\frac{1}{2}e^{-r^2}\right]_0^{\infty} \;=\; \pi. \end{aligned} \tag{8.65}$$

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 8.2.14. *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 8.2.15 (Cauchy Distribution)**.** *Consider*

$$p(x) \;=\; \frac{1}{\pi}\frac{1}{1+x^2}. \tag{8.66}$$

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,

$$\lim_{A\to\infty}\int_{-A}^{A} xp(x)dx \;=\; 0, \quad \lim_{A\to\infty}\int_{-A}^{2A} xp(x)dx \;=\; \infty. \tag{8.67}$$

Regardless, $p(x)$ has infinite variance. We shall see the Cauchy distribution again in Chapter 15; see also Exercises 3.3.28 and 3.3.29.

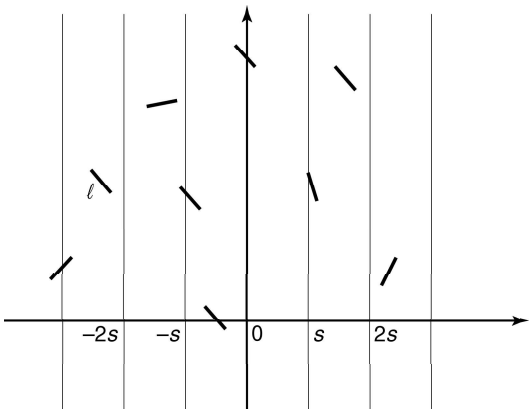

Figure 8.1 Buffon's needle

Exercise 8.2.16. *Prove the Cauchy distribution is a probability distribution by showing*

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx \ = \ 1. \tag{8.68}$$

Show the variance is infinite. See also Exercise 3.3.29.

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 15 and 16; see in particular §15.3.2.

8.2.4 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 $x = 0, \pm s, \pm 2s, \ldots$. 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 8.1. 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

$$p = \int_{\theta=0}^{2\pi} \int_{x=-\ell\cdot|\cos\theta|}^{\ell\cdot|\cos\theta|} \frac{dx d\theta}{s \cdot 2\pi} = 2\int_{\theta=0}^{2\pi} \frac{\ell \cdot |\cos\theta|}{s} \frac{d\theta}{2\pi} = \frac{2\ell}{\pi s}. \quad (8.69)$$

Exercise 8.2.17. *Show*

$$\frac{1}{2\pi}\int_0^{2\pi} |\cos\theta| d\theta = \frac{2}{\pi}. \quad (8.70)$$

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

$$A = \begin{cases} 1 & \text{with probability } \frac{2\ell}{\pi s} \\ 0 & \text{with probability } 1 - \frac{2\ell}{\pi s}. \end{cases} \quad (8.71)$$

If we were to throw N rods independently, since the expected value of a sum is the sum of the expected values (Lemma 8.1.45), 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

$$I \approx N \cdot \frac{2\ell}{\pi s} \quad \text{which implies} \quad \pi \approx \frac{N}{I} \cdot \frac{2\ell}{s}. \quad (8.72)$$

The above is an *experimental* formula for π!

Exercise 8.2.18. *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 §8.2.1, 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.*

8.3 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 8.1.45, 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 §8.4).

Exercise 8.3.1. *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 10 for the answer.

Exercise 8.3.2 (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 §15.3.2); it is the probability of observing a number at most x.*

Remark 8.3.3. 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 8.1.49, 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.

8.4 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 §8.4.2 and sketch the proof in general in §11.5. 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 9.

8.4.1 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 §8.3, 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 8.1.45 $\mathbb{E}[S_N] = N\mu$. 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 8.4.1 (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 $N \to \infty$*

$$\text{Prob}(S_N \in [\alpha, \beta]) \ \sim \ \frac{1}{\sqrt{2\pi\sigma^2 N}} \int_\alpha^\beta e^{-(t-\mu N)^2/2\sigma^2 N} dt. \tag{8.73}$$

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

$$\lim_{N\to\infty} \text{Prob}\left(\frac{S_N - \mu N}{\sqrt{\sigma^2 N}} \in [a, b]\right) \ = \ \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt. \tag{8.74}$$

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 12.7.8 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 §11.5 and [Fe].

Exercise 8.4.2. *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 8.4.3. *Redo Exercise 8.1.63 using the Central Limit Theorem and compare the two bounds.*

Exercise 8.4.4. *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 9 for more details and Exercise 12.7.8 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 3.3.1 and §4.4, and will see more in our investigations of writing integers as the sum of primes (see §13.3.2).

8.4.2 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 8.4.5 (Stirling's Formula)**.** *For n large,*

$$n! \;=\; n^n e^{-n}\sqrt{2\pi n}\,(1 + O(1/n))\,. \tag{8.75}$$

For a proof, see [WW]. We show (8.75) 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

$$\log n! \;=\; \sum_{k=1}^{n} \log k \;\approx\; \int_1^n \log t dt \;=\; (t\log t - t)|_1^n. \tag{8.76}$$

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

$$\begin{aligned}\binom{2N}{N+k} \;&\approx\; \frac{(2N)^{2N}}{(N+k)^{N+k}(N-k)^{N-k}}\sqrt{\frac{N}{\pi(N+k)(N-k)}} \\ &=\; \frac{2^{2N}}{\sqrt{\pi N}}\frac{1}{(1+\frac{k}{N})^{N+\frac{1}{2}+k}(1-\frac{k}{N})^{N+\frac{1}{2}-k}}.\end{aligned} \tag{8.77}$$

We would like to use $\left(1+\frac{w}{N}\right)^N \approx e^w$ from §5.4; 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

$$\left(1+\frac{k}{N}\right)^k \;=\; \left(1+\frac{k}{N}\right)^{N\cdot\frac{k}{N}} \;\to\; e^{k^2/N}; \tag{8.78}$$

similarly, $(1-\frac{k}{N})^{-k} \to e^{k^2/N}$. Thus we would claim (*and we shall see later in Lemma 8.4.6 that this claim is in error!*) that

$$\left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \;\to\; e^{2k^2/N}. \tag{8.79}$$

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 §5.4, 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 8.1.62) 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 8.4.8 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 8.4.6. *For any $\epsilon \le \frac{1}{9}$, for $N\to\infty$ with $k \ll N^{\frac{1}{2}+\epsilon}$, we have*

$$\left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k} \to e^{k^2/N}e^{O(N^{-1/6})}. \tag{8.80}$$

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

$$\log(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}x^n}{n}. \tag{8.81}$$

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

$$\begin{aligned}
P_{k,N} &= \left(1+\frac{k}{N}\right)^{N+\frac{1}{2}+k}\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k}\\
\log P_{k,N} &= \left(N+\frac{1}{2}+k\right)\log\left(1+\frac{k}{N}\right)+\left(N+\frac{1}{2}-k\right)\log\left(1-\frac{k}{N}\right)^{N+\frac{1}{2}-k}\\
&= \left(N+\frac{1}{2}+k\right)\left(\frac{k}{N}-\frac{k^2}{2N^2}+O\left(\frac{k^3}{N^3}\right)\right)\\
&\quad + \left(N+\frac{1}{2}-k\right)\left(-\frac{k}{N}-\frac{k^2}{2N^2}+O\left(\frac{k^3}{N^3}\right)\right)\\
&= \frac{2k^2}{N}-2\left(N+\frac{1}{2}\right)\frac{k^2}{2N^2}+O\left(\frac{k^3}{N^2}+\frac{k^4}{N^3}\right)\\
&= \frac{k^2}{N}+O\left(\frac{k^2}{N^2}+\frac{k^3}{N^2}+\frac{k^4}{N^3}\right).
\end{aligned} \tag{8.82}$$

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

$$P_{k,N} = e^{k^2/N}e^{O\left(N^{-1/6}\right)}, \tag{8.83}$$

which completes the proof. □

Combining Lemma 8.4.6 with (8.77) yields

$$\binom{2N}{N+k}\frac{1}{2^{2N}} \approx \frac{1}{\sqrt{\pi N}}\, e^{-k^2/N}. \tag{8.84}$$

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

$$\text{Prob}(S_{2N} = 2k) \;=\; \binom{2N}{N+k}\frac{1}{2^{2N}} \;\approx\; \frac{2}{\sqrt{2\pi\cdot(2N)}}\, e^{-(2k)^2/2(2N)}. \tag{8.85}$$

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 8.4.7. *Use the integral test to bound the error in* (8.76)*, and then use that to bound the error in the estimate of* $n!$.

Exercise 8.4.8. *Prove the standard deviation of* S_{2N} *is* $\sqrt{2N}$*. Use this and Chebyshev's inequality (Exercise 8.1.62) to prove*

$$\text{Prob}(|S_{2N}| \geq N^{\epsilon}\cdot\sqrt{2N}) \;\leq\; \frac{1}{N^{2\epsilon}}, \tag{8.86}$$

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

Exercise 8.4.9. *Prove* (8.81)*.*

Exercise 8.4.10. *Can you generalize the above arguments to handle the case when* $p \neq \frac{1}{2}$.

Chapter Nine

Applications of Probability: Benford's Law and Hypothesis Testing

The Gauss-Kuzmin Theorem (Theorem 10.3.1) 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 12), logarithmic probabilities (similar to the Gauss-Kuzmin probabilities in Chapter 10), and Poisson Summation (Chapter 11), as well as many of the common problems in statistical testing (such as non-independent events and multiple comparisons).

9.1 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,

$$\lim_{N\to\infty} \frac{\#\{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\}}{N} \;=\; \log_b\left(1+\frac{1}{j}\right). \tag{9.1}$$

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

$$\sum_{j=1}^{b-1} \log_b\left(1+\frac{1}{j}\right) \;=\; \log_b \prod_{j=1}^{b-1}\left(1+\frac{1}{j}\right) \;=\; \log_b \prod_{j=1}^{b-1} \frac{j+1}{j} \;=\; \log_b b \;=\; 1. \tag{9.2}$$

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

$$a_{n+1} \;=\; \begin{cases} 3a_n+1 & \text{if } a_n \text{ is odd} \\ a_n/2 & \text{if } a_n \text{ is even.} \end{cases} \tag{9.3}$$

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

$$\begin{aligned} 13 \longrightarrow 40 \longrightarrow 20 \longrightarrow 10 \longrightarrow 5 \longrightarrow 16 \longrightarrow 8 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \\ \longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \longrightarrow 4 \longrightarrow 2 \longrightarrow 1 \cdots. \end{aligned} \tag{9.4}$$

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

$$a_{n+1} \;=\; \frac{3a_n+1}{2^k}, \tag{9.5}$$

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 9.1.1. *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 9.1.2. *Below we use the definition of the* $3x+1$ *map from* (9.5)*. 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* $a_k = 1$*?*

9.2 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 9.2.1. *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

$$b^u = u_k b^k + u_{k-1} b^{k-1} + \cdots + u_0 + u_{-1} b^{-1} + \cdots, \tag{9.6}$$

then

$$\begin{aligned} b^v &= b^{u+m} \\ &= b^u \cdot b^m \\ &= (u_k b^k + u_{k-1} b^{k-1} + \cdots + u_0 + u_{-1} b^{-1} + \cdots) b^m \\ &= u_k b^{k+m} + \cdots + u_0 b^m + u_{-1} b^{m-1} + \cdots. \end{aligned} \tag{9.7}$$

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

Exercise 9.2.2. *Prove the other direction of the if and only if.*

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

$$b^{p_j} = j \quad \text{or equivalently} \quad p_j = \log_b j. \tag{9.8}$$

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

$$I_j^{(b)} = [p_j, p_{j+1}) \subset [0, 1). \tag{9.9}$$

Lemma 9.2.3. *The first digit of b^y base b is j if and only if y mod $1 \in I_j^{(b)}$.*

Proof. By Lemma 9.2.1 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 12.1.4) are Benford.

Theorem 9.2.4. *If $y_n = \log_b x_n$ is equidistributed mod 1 then x_n is Benford (base b).*

Proof. By Lemma 9.2.3,

$$\begin{aligned} &\{n \le N : y_n \bmod 1 \in [\log_b j, \log_b (j+1))\} \\ &\quad = \{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\}. \end{aligned} \tag{9.10}$$

Therefore

$$\begin{aligned} &\lim_{N \to \infty} \frac{\#\{n \le N : y_n \bmod 1 \in [\log_b j, \log_b (j+1))\}}{N} \\ &\quad = \lim_{N \to \infty} \frac{\#\{n \le N : \text{first digit of } x_n \text{ in base } b \text{ is } j\}}{N}. \end{aligned} \tag{9.11}$$

If y_n is equidistributed, then the left side of (9.11) is $\log_b \left(1 + \frac{1}{j}\right)$ which implies x_n is Benford base b. □

Remark 9.2.5. 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 9.2.4 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 12.3.2 we prove that for $\alpha \notin \mathbb{Q}$ the fractional parts of $n\alpha$ are equidistributed modulo 1. From this and Theorem 9.2.4, it immediately follows that geometric series are Benford (modulo the irrationality condition):

Theorem 9.2.6. *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 12.3.2 the fractional parts of y_n are equidistributed. Exponentiating by b, we obtain that x_n is Benford (base b) by Theorem 9.2.4. □

Theorem 9.2.6 implies that 2^n is Benford base 10, but not surprisingly that it is not Benford base 2.

Exercise 9.2.7. *Do the first digits of e^n follow Benford's Law? What about $e^n + e^{-n}$?*

9.3 RECURRENCE RELATIONS AND BENFORD'S LAW

We show many sequences defined by recurrence relations are Benford. For more on recurrence relations, see Exercise 7.3.9. The interested reader should see [BrDu, NS] for more on the subject.

9.3.1 Recurrence Preliminaries

We consider recurrence relations of length k:

$$a_{n+k} = c_1 a_{n+k-1} + \cdots + c_k a_n, \tag{9.12}$$

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

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - \cdots - c_{k-1} r - c_k = 0 \tag{9.13}$$

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

$$a_n = u_1 \lambda_1^n + \cdots + u_k \lambda_k^n, \tag{9.14}$$

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

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 7.3.11). If $|\lambda_1| = 1$, we do not expect the first digit of a_n to be Benford (base b). For example, if we consider

$$a_n = 2a_{n-1} - a_{n-2} \tag{9.15}$$

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.

9.3.2 Recurrence Relations Are Benford

Theorem 9.3.1. *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 9.2.4 it suffices to show y_n is equidistributed mod 1. We have

$$\begin{aligned} x_n &= u_1\lambda_1^n + \cdots + u_n\lambda_k^n \\ x_n &= u_1\lambda_1^n \left[1 + O\left(\frac{ku\lambda_2^n}{\lambda_1^n}\right)\right], \end{aligned} \tag{9.16}$$

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

$$\beta = \frac{(ku)^{1/n_0}}{\lambda_1^{\epsilon}} \frac{|\lambda_2|}{\lambda_1^{1-\epsilon}} < 1, \tag{9.17}$$

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

$$\begin{aligned} y_n &= \log_b x_n \\ &= \log_b(u_1\lambda_1^n) + O\left(\log_b(1+\beta^n)\right) \\ &= n\log_b \lambda_1 + \log_b u_1 + O(\beta^n), \end{aligned} \tag{9.18}$$

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, $j \in \{1, \ldots, b-1\}$. This means

$$n\log_b \lambda_1 + \log_b u_1 \in I_j^{(b)} = [p_{j-1}, p_j). \tag{9.19}$$

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 9.3.2. *Weaken the conditions of Theorem 9.3.1 as much as possible. What if several roots equal λ_1? What does a general solution to* (9.12) *look like now? What if λ_1 is negative? Can anything be said if there are complex roots?*

Exercise[(hr)] 9.3.3. *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)] 9.3.4. *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.*

9.4 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 8.4.1), 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 9.4.1.

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$^{(hr)}$ 9.4.1. *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 $\mathrm{Prob}(k = \ell) = 2^{-\ell}$.*

9.4.1 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 9.4.2.

$$\frac{2}{\sqrt{2\pi\sigma^2}} \int_{\sigma^{1+\delta}}^{\infty} e^{-x^2/2\sigma^2} dx \ll e^{-\sigma^{2\delta}/2}. \tag{9.20}$$

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

$$I = \frac{2}{\sqrt{2\pi\sigma^2}} \int_{\sigma^\delta/\sqrt{2}}^{\infty} e^{-w^2} \cdot \sigma\sqrt{2} dw = \frac{2}{\sqrt{\pi}} \int_{\sigma^\delta/\sqrt{2}}^{\infty} e^{-w^2} dw. \tag{9.21}$$

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,

$$\begin{aligned} I &< 1 \cdot \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}+1}^{\infty} e^{-w^2} dw \\ &= \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-(u+1)^2} du \\ &= \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{\sqrt{\pi}} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-u^2} e^{-2u} e^{-1} du \\ &< \frac{2}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} + \frac{2}{e\sqrt{\pi}} e^{-\sigma^{2\delta}/2} \int_{\frac{\sigma^\delta}{\sqrt{2}}}^{\infty} e^{-2u} du \\ &< \frac{2(e+1)}{\sqrt{\pi}} e^{-\sigma^{2\delta}/2} \\ &< 4e^{-\sigma^{2\delta}/2}. \end{aligned} \tag{9.22}$$

□

Exercise 9.4.3. *Prove a similar result for intervals of the form* $[-\sigma g(\sigma), \sigma g(\sigma)]$ *where* $g(\sigma)$ *is a positive increasing function and* $\lim_{\sigma\to\infty} g(\sigma) = +\infty$.

9.4.2 Geometric Brownian Motions Are Benford

We investigate the distribution of digits of processes that are Geometric Brownian Motions. By Theorem 9.2.4 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

$$p_N(x) \;=\; \frac{e^{-\pi x^2/N}}{\sqrt{N}}. \tag{9.23}$$

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 9.4.4. *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

$$\int_{\substack{x=-\infty \\ x \bmod 1 \in [a,b]}}^{\infty} p_N(x)dx \;=\; \frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx. \tag{9.24}$$

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

$$e^{-\pi(x+n)^2/N} \;=\; e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N}\, e^{-n^2/N}\right). \tag{9.25}$$

We claim that in (9.24) it is sufficient to restrict the summation to $|n| \le N^{5/4}$. The proof is immediate from Lemma 9.4.2: 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 §11.4.2 we prove the Poisson Summation formula, which in this case yields

$$\frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} e^{-\pi n^2/N} \;=\; \sum_{n\in\mathbb{Z}} e^{-\pi n^2 N}. \tag{9.26}$$

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 (9.26) 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

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{|n|\le N^{5/4}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx \\
&\quad= \frac{1}{\sqrt{N}} \sum_{|n|\le N^{5/4}} \int_{x=a}^{b} \left[e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N} e^{-n^2/N}\right)\right] dx \\
&\quad= \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \sum_{n=0}^{N^{5/4}} \frac{n+1}{\sqrt{N}} e^{-\pi(n/\sqrt{N})^2}\right) \\
&\quad= \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \int_{w=0}^{N^{3/4}} (w+1) e^{-\pi w^2} \sqrt{N} dw\right) \\
&\quad= \frac{b-a}{\sqrt{N}} \sum_{|n|\le N^{5/4}} e^{-\pi n^2/N} + O\left(N^{-1/2}\right).
\end{aligned} \tag{9.27}$$

By Lemma 9.4.2 we can extend all sums to $n \in \mathbb{Z}$ in (9.27) with negligible error. We now apply Poisson Summation and find that up to lower order terms,

$$\frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx \;\approx\; (b-a) \cdot \sum_{n\in\mathbb{Z}} e^{-\pi n^2 N}. \tag{9.28}$$

For $n = 0$ the right hand side of (9.28) is $b-a$. For all other n, we trivially estimate the sum:

$$\sum_{n\neq 0} e^{-\pi n^2 N} \;\leq\; 2\sum_{n\geq 1} e^{-\pi n N} \;\leq\; \frac{2e^{-\pi N}}{1-e^{-\pi N}}, \tag{9.29}$$

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 9.4.5. 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 5.5.1; see [LaSo] for additional results on Benford behavior of the $3x+1$ problem.

9.5 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.

9.5.1 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.

9.5.2 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 10.3.1 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 §8.3). Consider the claim that a process has probability p of success. We have N independent Bernoulli trials (see §8.2.1). 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 §8.4) we expect S_N to have a Gaussian distribution with mean pN and variance pqN (see Exercise 8.2.1 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:

$$\frac{1}{\sqrt{2\pi pqN}} \int_{|s-pN|\geq a} e^{-(s-pN)^2/2pqN} ds. \tag{9.30}$$

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 9.5.1. *Prove the above statement about the distribution of z.*

Let

$$I(a) = \frac{1}{\sqrt{2\pi}} \int_{|z| \geq a} e^{-z^2/2} dz, \tag{9.31}$$

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 9.5.2. 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.

9.5.3 Digits of the $3x + 1$ Problem

Consider again the $3x + 1$ problem. Choose a large integer a_0, and look at the iterates: $a_1, a_2, a_3, \ldots$. We study how often the first digit of terms in the sequence equal $d \in \{1, \ldots, 9\}$. 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 9.5.3. *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 $a_0 = 333\ldots333$ 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 9.5.4 (Important)**.** One must be very careful when analyzing all the digits. Once we know how many digits are in $\{1,\ldots,8\}$, 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 9.5.5). 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 9.5.5. *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 9.5.6. 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,

$$3 \cdot [0, \dots, 0, 0, 5, 6, 7] \ = \ [0, \dots, 0, 1, 7, 0, 1]. \tag{9.32}$$

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 9.5.7. *Consider the first digits of the $3x + 1$ problem (defined as in* (9.3)*) 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 9.5.8 (Recommended)**.** *Here is another variant of the $3x + 1$ problem:*

$$a_{n+1} \ = \ \begin{cases} 3a_n + 1 & \text{if } a_n \text{ is odd} \\ a_n/2^k & \text{if } a_n \text{ is even and } 2^k || a_n; \end{cases} \tag{9.33}$$

$2^k || a_n$ 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

$$a_{n+1} \ = \ \frac{3a_n + 1}{2^k}, \quad 2^k || a_n. \tag{9.34}$$

9.5.4 Digits of Continued Fractions

Let us test the hypothesis that the digits of algebraic numbers are given by the Gauss-Kuzmin Theorem (Theorem 10.3.1). 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 10 for more detailed experiments on algebraic numbers and the Gauss-Kuzmin Theorem.

9.6 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.