BROWN UNIVERSITY
Math 1610 Probability Notes
Samuel S. Watson
Last updated: December 18, 2015

Please do not hesitate to notify me about any mistakes you find in these notes. My advice is to refresh your local copy of this document frequently, as I will be updating it throughout the semester.

1 Probability Spaces

We model random phenomena with a *probability space*, which consists of an arbitrary set Ω, a collection[1] $\mathcal{F}$ of subsets of Ω, and a map $P : \mathcal{F} \to [0,1]$, where $\mathcal{F}$ and P satisfy certain conditions detailed below. An element $\omega \in \Omega$ is called an *outcome*, an element $E \in \mathcal{F}$ is called an *event*, and we interpret $P(E)$ as the probability of the event E. To connect this setup to the way you usually think about probability, regard ω as having been randomly selected from Ω in such a way that, for each event E, the probability that ω is in E (in other words, that E happens) is equal to $P(E)$.

If E and F are events, then the event "E and F" corresponds to the $\{\omega : \omega \in E \text{ and } \omega \in F\}$, abbreviated as $E \cap F$. Similarly, $E \cup F$ is the event that E happens or F happens, and $\Omega \setminus E$ is the event that E does not happen. We refer to $\Omega \setminus E$ as the *complement* of E, and sometimes denote[2] it by E^c.

To ensure that we can perform these basic operations, we require that $\mathcal{F}$ is *closed* under them. In other words, $\Omega \setminus E$ must be an event whenever E is an event (that is, $\Omega \setminus E \in \mathcal{F}$ whenever $E \in \mathcal{F}$). For the union, we require not only that $E \cup F$ is an event whenever E or F are events, but the following countably infinite version: if $(E_i)_{i=1}^{\infty}$ is a sequence of events, then $\bigcup_{i=1}^{\infty} E_i$ is also an event. The final constraint on $\mathcal{F}$ is that that Ω itself is an event. Solve the following exercise to see why we omitted intersections from this list of requirements.

Exercise 1.1. Show that if $\mathcal{F}$ is closed under countable unions and complements, then it is also closed under countable intersections.

If $\mathcal{F}$ satisfies these conditions, we say that $\mathcal{F}$ is a *σ-algebra* (read "sigma algebra").

Similarly, P must satisfy certain properties if we want to interpret its values as probabilities. First, we require $P(E \cup F) = P(E) + P(F)$ whenever $E \cap F = \emptyset$ (in other words, if

[1] We did not talk about $\mathcal{F}$ in class, as it is not discussed in the book. I have chosen to include it here for completeness.

[2] Your book uses $\tilde{E}$.

E and F cannot both happen, then the probability that E or F happens is the sum of the probabilities of E and F). We call this property of P *additivity*. Furthermore, we require the countable version[3] $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$ whenever $(E_i)_{i=1}^{\infty}$ is pairwise disjoint (that is, $E_i \cap E_j = \emptyset$ whenever $i \neq j$). Finally, we require $P(\Omega) = 1$. Other properties we would want P to have will follow from these (see Exercise 1.2). A map $P : \mathcal{F} \to [0,1]$ satisfying these conditions is called a *probability measure* on $\mathcal{F}$. For $E \in \mathcal{F}$, the value of $P(E)$ may be referred to as the *measure* of E.

Exercise 1.2. Show that if P is a probability measure on $\mathcal{F}$, then

1. $P(E) \leq P(F)$ whenever $E \subset F$,
2. $P(E^c) = 1 - P(E)$ for all $E \in \mathcal{F}$,
3. $P(\emptyset) = 0$.

In summary, if Ω is a set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and P is a probability measure on $\mathcal{F}$, we say that $(\Omega, \mathcal{F}, P)$ is a probability space. We often wish to ignore the role of $\mathcal{F}$, in which case we may refer to P as a probability measure on Ω.

1.1 Discrete probability spaces

If Ω is finite or countably infinite, then we typically take $\mathcal{F}$ to be the set of all subsets of Ω, and we describe P by considering its values on all singleton sets: we define

$$m(\omega) = P(\{\omega\}),$$

the probability of the outcome ω, and we call m a *probability mass function*[4], abbreviated pmf. The reason for this name is that we think of probability as one unit of "probability mass" that is to be shared among the outcomes $\omega \in \Omega$. Note that the probability of any event E can be recovered from the values of m using additivity:

$$P(E) = P\left(\bigcup_{\omega \in E} \{\omega\}\right) = \sum_{\omega \in E} m(\omega).$$

Conversely, if $m : \Omega \to [0,1]$ has the property that $\sum_{\omega \in \Omega} m(\omega) = 1$, then defining $P(E) = \sum_{\omega \in E} m(\omega)$ does indeed give a probability measure (exercise).

Example 1.3. Consider two coin flips. We may take as our set of possible outcomes

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

our σ-algebra is the set of all subsets of Ω, and finally $m(\omega) = 1/4$ for each outcome $\omega \in \Omega$ (this is called the uniform distribution on Ω).

[3] We might like to drop the restriction that the collection of events is countable in this property, but that would be too much to ask. See further discussion in Section 1.2.

[4] Your book calls it a probability distribution function.

A *random variable* X is a function from Ω to $\mathbb{R}$. For instance, consider the function X which maps an element ω from Example 1.3 to the number of heads in ω times 3. You can think of X as describing a payout for each outcome $\omega \in \Omega$. We frequently suppress the role of ω in our notation, making frequent use of abbreviations such as $P(X = 3)$ for $P(\{\omega \in \Omega : X(\omega) = 3\})$.

Example 1.4. In the context of Example 1.3, define $X(\omega)$ to be three times the number of heads in ω, and find $P(X = 3)$.

Solution. The event $\{X = 3\}$ includes the outcomes HT and TH, but not TT or HH. Therefore, $P(X = 3) = P(\{\text{HT}, \text{TH}\}) = 1/4 + 1/4 = 1/2$. □

The *law* or *distribution* of a random variable X is the probability measure Q on $\mathbb{R}$ (or on a subset[5] of $\mathbb{R}$) defined by $Q(A) = P(X \in A)$. In Example 1.3, for instance, Q is the measure on $\{0, 3, 6\}$ with mass distribution $(m(0), m(3), m(6)) = (1/4, 1/2, 1/4)$. We emphasize that P is a probability measure on $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{HH}\}$ while Q is a measure on $\{0, 3, 6\} \subset \mathbb{R}$. We use the notation $X \sim Q$ to mean that the law of X is Q.

If two random variables X and Y are defined on the same probability space, then the *joint distribution* of X and Y is defined to the measure on $\mathbb{R}^2$ which maps each set $A \subset \mathbb{R}^2$ to the probability that the ordered pair (X, Y) is in A. If we have in mind a pair of random variables X and Y and want to discuss the law of X while emphasizing that we're doing so without regard for Y, we call the law of X the *marginal law* or *marginal distribution* of X.

Example 1.5. A common way to specify a discrete probability measure is to draw a *tree diagram*—see the discussion page 24 in your book.

1.2 Continuous probability spaces

If Ω is uncountably infinite, then it is possible that $P(\{\omega\}) = 0$ for all $\omega \in \Omega$, even though $P(\Omega) = 1$. For example, consider $\Omega = [0, 1]$ and $P(E) = \text{length}(E)$. If we check the requirements, it makes sense, and indeed is true[6], that P is a probability measure, and of course it has the property that the measure of each singleton set is zero. Therefore, in this context we cannot use a probability mass function as we did in the discrete setting.

In general, rigorously defining a probability measure for uncountable Ω is pretty technical and beyond the scope of this course, so we will avail ourselves of the length measure and

[5] If Q is a probability measure on a subset $B \subset \mathbb{R}$, we can think of Q as a probability measure on $\mathbb{R}$ by assigning zero probability mass to $\mathbb{R} \setminus B$; in other words, we define $Q(A)$ to be $Q(A \cap B)$ for all $A \subset \mathbb{R}$. For this reason, we may always regard Q as a probability measure on $\mathbb{R}$, and it is convenient to do so.

[6] With the length function suitably extended to a σ-algebra $\mathcal{F}$ called the *Borel* σ-algebra. When you take a measure theory course, you will learn that $\mathcal{F}$ does not include *all* subsets of Ω, it is a very rich class of sets that includes all the subsets of $[0, 1]$ you're ever likely to encounter.

its cousins *area* in $\mathbb{R}^2$ and *volume* in three dimensions and higher. For this reason, we will insist in this course that if Ω is uncountable, it is a subset of $\mathbb{R}^n$ for some positive integer n.

Suppose that $\Omega \subset \mathbb{R}^n$, and suppose that $f : \Omega \to [0, \infty)$ has the property that[7] $\int_\Omega f dV = 1$. We call f a *probability density function*, abbreviated pdf, and we define

$$P(E) = \int_E f \, dV$$

for all $E \subset \Omega$. Then P is indeed a probability measure[8]. The reason f is called a density is that $f(\omega)$ equals the limit as $\epsilon \to 0$ of the probability of the ball of radius ϵ around ω divided by the volume (or area/length) of that ball. In other words, if we have in mind the analogy between probability and mass, then f corresponds to density in the physics sense: mass divided by volume.

Example 1.6. Consider the probability space with $\Omega = [0, 1]$ and probability measure given by the density $f(x) = 2x$ for $x \in [0, 1]$. Find $P([1/2, 1])$.

Solution. We calculate $P([1/2, 1]) = \int_{1/2}^1 2x \, dx = 3/4$. □

If f is constant on Ω, then we call f the *uniform measure* on Ω. Note that this requires that Ω have finite volume.

1.3 Choosing coordinates for continuous probability spaces

A *chord* of a circle is a line segment whose endpoints are on the circle. A random chord is a chord which is selected randomly in some way. Mathematically, we describe a random chord by defining Ω to be the set of all chords of the unit circle and defining some probability measure P on Ω. Note that Ω is uncountable and is not a subset of any Euclidean space $\mathbb{R}^n$ (an element of Ω is a chord, not a point in $\mathbb{R}^n$). Since our only tools for constructing probability measures on uncountable spaces are for Euclidean spaces, we must transfer the problem to a Euclidean space by finding a bijection between Ω and some set $\widetilde{\Omega} \subset \mathbb{R}^n$.

For example, consider the map φ from Ω to the triangle

$$\widetilde{\Omega} = \{(x, y) \, : \, 0 \leq x < y < 2\pi\}$$

[7] We'll denote the volume differential in $\mathbb{R}^n$ by dV, but note that if $n = 2$ it would be more standard to call it the area differential and write it as dA or $dx \, dy$, and if $n = 1$ it would just be dx, the length differential.

[8] A complete proof is beyond the scope of the course, but note that additivity follows from basic properties of the integral.

which associates with each chord C, whose endpoints have angles $0 \leq \theta_1 < \theta_2 < 2\pi$ with respect to the positive x-axis, the ordered pair (θ_1, θ_2). We may then equip Ω with a probability measure P by choosing a probability measure $\widetilde{P}$ on $\widetilde{\Omega}$ and defining

$$P(E) = \widetilde{P}(\varphi(E)).$$

One particularly natural choice for $\widetilde{P}$ is the uniform measure on $\widetilde{\Omega}$.

We call φ a *choice of coordinates* for Ω, and there are other natural choices of coordinates for Ω. Some measures P obtained using obtained using these other choices of coordinates may be different, and indeed your book describes some that are in fact different. See page 47 in your book for further discussion of this phenomenon, called *Bertrand's paradox*[9].

1.4 Cumulative distribution functions

If $\Omega \subset \mathbb{R}$, then we can describe P by its *cumulative distribution function*, abbreviated cdf, which is defined to be the function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = P((-\infty, x]),$$

If P has density function f, then

$$F(x) = \int_{-\infty}^{x} f(t)\, dt.$$

Note that, by the fundamental theorem of calculus, F and f are related by

$$F'(x) = f(x).$$

One nice thing about F is that it unifies discrete and continuous random variables. If the law of a random variable X has density f, then $F(x)$ increases continuously (actually, even smoother: differentiably) from 0 to 1 as x goes from $-\infty$ to ∞. If, by contrast, X is a random variable that only takes on countably many values $\cdots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \cdots$, then F makes a discontinuous jump at each value x_i and is constant on each interval $[x_i, x_{i+1})$.

Cumulative distribution functions can be useful when reasoning about the distributions of random variables:

Example 1.7. Suppose that X is a uniform random variable on $[0, 1]$ and $Y = X^2$ (this means that for all $\omega \in \Omega$, we have $Y(\omega) = (X(\omega))^2$). Find the cdf of X and the cdf of Y.

[9] As far as I can tell, the term "paradox" here is used because of the divergence between the aforementioned facts and the incorrect intuition that there should be one canonical probability measure on the set of chords (that is, a probability measure that would be the same for any reasonable choice of coordinates, using uniform measure on $\widetilde{\Omega}$).

Solution. The cdf F_X of X is given by

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

To find the cdf F_Y of Y, we note that $Y \leq y$ if and only if $X^2 \leq y$, which is true if and only if $X \leq \sqrt{y}$. Thus

$$F_Y(y) = P(X \leq \sqrt{y}) = \begin{cases} 0 & \text{if } y \leq 0 \\ \sqrt{y} & \text{if } 0 \leq y \leq 1 \\ 1 & \text{if } 1 \leq y. \end{cases}$$ □

Exercise 1.8. Find the pdfs of X and Y. More generally, derive a formula for the pdf of $Y = g(X)$ in terms of the pdf of X, where g is an increasing differentiable function.

2 Counting

2.1 Permutations

A *permutation* σ is a bijective map from a finite set A to itself. It is convenient to take A to be $\{1, 2, 3, \ldots, n\}$, where n is a positive integer, and we write the permutation in the form $\sigma(1)\sigma(2)\cdots\sigma(n)$ (we're just listing the values in order—one should not interpret the juxaposition as multiplication). For example, the permutations of $\{1, 2\}$ are the identity permutation 12 and the permutation 21 which maps 1 to 2 and 2 to 1.

Theorem 2.1. The number of permutations on $\{1, 2, \ldots, n\}$ is $n!$.

Proof. We may form a permutation by choosing any of the values from 1 to n to be the value of $\sigma(1)$. Then, regardless of what we chose for $\sigma(1)$, we can choose any of the remaining $n-1$ values for $\sigma(2)$. By the *rule of product*[10], there are $n(n-1)$ ways of choosing $\sigma(1)$ and $\sigma(2)$. Continuing in this way, we get $n(n-1)(n-2)\cdots\cdots 2\cdot 1 = n!$ ways of choosing the whole collection of values $\sigma(1), \ldots \sigma(n)$. □

We often want to approximate an expression involving factorials, and for that the following approximation is very useful. We say that $a_n \sim b_n$, read as "a_n is asymptotic to b_n" if $a_n/b_n \to 1$ as $n \to \infty$.

Theorem 2.2. We have $n! \sim (n/e)^n\sqrt{2\pi n}$.

[10]The rule of product says that if there are are A ways to make on choice and B ways to make a second choice, then there are AB ways to make the pair of choices

We will not prove Stirling's formula. We note that there exist stronger versions of this approximation which include more terms and are more accurate.

To generate a random permutation, we can choose $\sigma(1)$ uniformly at random from $\{1, 2, \ldots, n\}$, then choose $\sigma(2)$ uniformly at random from $\{1, 2, \ldots, n\} \setminus \{\sigma(1)\}$, and so on.

2.2 Combinations

If n and k are positive integers, we define $\binom{n}{k}$ (read as "n choose k") to be the number of ways to choose a subset of size k from a set of size n. We first count the number of ordered k-tuples of elements of $\{1, 2, \ldots, n\}$ by choosing an arbitrary element of $\{1, 2, \ldots, n\}$ for the first value, any other element for the second value, and so on until we have chosen all k values for our k-tuple. Thus there are

$$n(n-1)(n-2)\cdots(n-k+1)$$

ordered k-tuples (note that we stop at $n-k+1$ so that there are k total factors). We abbreviate this quantity $(n)_k$ and read it as "n down k".

Note that each set of k elements corresponds to $k!$ different k-tuples. Thus the count $(n)_k$ represents an overcount of the sets by a factor of $k!$. Thus we may conclude that

$$\binom{n}{k} = \frac{(n)_k}{k!} = \frac{n!}{k!(n-k)!},$$

where in the second equality we have rewritten $(n)_k$ as $n!/(n-k)!$. We can prove many properties of binomial coefficients using the counting definition:

Proposition 2.3. We have

$$\binom{n}{j} = \binom{n-1}{j} + \binom{n-1}{j-1}.$$

Proof. The first term on the right hand side equals the number of ways to choose a subset of $\{1, 2, \ldots, n\}$ of size j that does not include n, and the second term equals the number of ways to choose a subset of size j that *does* include n (since that boils down to choosing a subset of $\{1, 2, \ldots, n-1\}$ of size $j-1$). Since a subset must either include n or not, the right-hand side equals the number of ways to choose a subset of size j from $\{1, 2, \ldots, n\}$, which in turn is equal to the left-hand side. □

3 Conditional probability

3.1 Discrete conditional probability

Suppose that (Ω, P) is a discrete probability space modeling a random experiment, and suppose $E \subset \Omega$. After the experiment has been performed, but before we see the outcome, a trustworthy party who *can* see the outcome ω tells us that $\omega \in E$, where E is some subset of Ω. What probability measure should be used to model this new situation? Clearly we should assign zero mass to the complement of E, since we know that ω is in E. And the mass assigned to each element of E should be proportional to its original mass. This leads us to define the *conditional probability mass function*, written as $m(\cdot|E)$, by

$$m(\omega|E) = m(\omega)/P(E),$$

for $\omega \in E$, and $m(\omega|E) = 0$ otherwise. The constant factor $1/P(E)$ was chosen so that $m(\cdot|E)$ has total mass 1; note that we must assume $P(E) > 0$. For any event F, we can sum over $\omega \in F$ to find that

$$P(F|E) = P(F \cap E)/P(E). \tag{3.1}$$

This is the fundamental equation defining conditional probability.

See Example 4.5 on page 135 in your book for a simple example of a conditional probability calculation.

3.2 Independence

The intuitive idea of independence is that two events E and F are independent if they have nothing to do with one another: in other words, if you know whether E occurs, that doesn't change the probability that F occurs. For example, two separate die rolls are independent, whereas the first die roll and the sum of the two die rolls are not (if you now the first die roll is 6, then the sum cannot be smaller than 7, for instance).

So, if $P(E) > 0$ we say that E and F are independent if $P(F|E) = P(F)$. If $P(E) = 0$, then we say E and F are independent for all F. Using (3.1), we can rewrite this relation in a more symmetric way: events E and F are independent if and only if

$$P(E \cap F) = P(E)P(F). \tag{3.2}$$

We often want to discuss independence of more than one event. For instance, if we flip 10 coins, then the relationship between the 10 flips is that no knowledge of any of the

flips would influence our assessment of the probabilities of any of the other flips. In other words, if A_i is the event that the ith flip comes up heads, then, for example,

$$P(E_1 \cap E_7 \cap E_8 \cap E_{10}) = P(E_1)P(E_7)P(E_8)P(E_{10}),$$

and similarly for every subset of $\{1, 2, \ldots, 10\}$. This leads us to the following definition:

We say that events $E_1, \ldots E_n$ are *mutually independent* (or just *independent*) if for all $k \leq n$ and all $1 \leq i_1 < \cdots i_k \leq n$, we have

$$P(E_{i_1} \cap \cdots \cap E_{i_k}) = P(E_{i_1}) \cdots P(E_{i_k}).$$

We will define independence for random variables using the definition of independence for events: we say that random variables $X_1, \ldots X_n$ (defined on the same probability space) are independent if for all sets $A_1 \subset \mathbb{R}$, $A_2 \subset \mathbb{R}$, ..., $A_n \subset \mathbb{R}$, the events $\{X \in A_1\}$, $\{X \in A_2\}$, ..., $\{X \in A_n\}$ are independent.

3.2.1 Product measure (the discrete case)

If X is a discrete random variable taking values in $\Omega_X \subset \mathbb{R}$ whose law has probability mass function m_X and Y is a discrete random variable taking values on $\Omega_Y \subset \mathbb{R}$ whose law has probability mass function m_Y, then we can always construct a probability space on which X and Y are independent. We set

$$\Omega = \{(\omega_1, \omega_2) \,:\, \omega_1 \in \Omega_X \text{ and } \omega_2 \in \Omega_Y\} = \Omega_X \times \Omega_Y$$

and $m((\omega_1, \omega_2)) = m_X(\omega_1)m_Y(\omega_2)$. Defining $X((\omega_1, \omega_2)) = \omega_1$ and $Y((\omega_1, \omega_2)) = \omega_2$, we find that X and Y have the prescribed laws, and X and Y are independent (left as an exercise). We call the measure associated with m the *product measure* associated with the measures m_X and m_Y. This construction (and its continuous analogue, discussed below) is the probability space analogue of the Cartesian product in set theory.

Defining product measure allows us to rephrase the definition of independence of random variables: the random variables $X_1, \ldots, X_n$ are independent if and only if their joint distribution is given by the product measure of the laws of the X_k's (exercise).

3.3 Continuous conditional probability

If (Ω, P) is a continuous probability space (so $\Omega \subset \mathbb{R}^n$) and P has probability density function f, we define the *conditional probability density* given $E \subset \Omega$ to be

$$f(x|E) = \begin{cases} f(x)/P(E) & \text{if } x \in E \\ 0 & \text{if } x \in \Omega \setminus E, \end{cases}$$

assuming $P(E) > 0$. With this definition, we obtain the same conditional probability formula

$$P(F|E) = \int_F f(x|E)\,dx = P(E \cap F)/P(E)$$

as in the discrete case (in this equation, $F \subset \Omega$ is an arbitrary event).

Exercise 3.1. Let $\Omega = [0, \infty)$, let $\lambda > 0$ be a constant, and let P be the probability measure whose density is given by $f(x) = \lambda e^{-\lambda x}$. Show that $P([t+s, \infty)|[t, \infty)) = P([s, \infty))$ for all $s, t > 0$.

This property is called the *memorylessness* of the exponential distribution $\lambda e^{-\lambda x}$.

3.3.1 Product measure: the continuous case

Given probability density functions f_X and f_Y on $\mathbb{R}$, it is always possible to construct independent random variables X and Y with probability density functions f_X and f_Y, respectively. We set $\Omega = \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, and we define

$$f(x, y) = f_X(x) f_Y(y).$$

Under the measure with probability density function f, the random variables $X((\omega_1, \omega_2)) = \omega_1$ and $Y((\omega_1, \omega_2)) = \omega_2$ are independent, while X and Y indeed have probability density functions f_X and f_Y, respectively (exercise). This construction generalizes to any finite number of random variables.

4 Miscellaneous topics

4.1 Binomial probabilities

Question 4.1. What is the probability of rolling exactly 18 sixes in 100 independent rolls of a fair die?

Proof. There are many ways to roll 18 sixes. We could roll 18 sixes followed by 82 non-sixes, and the probability of that is

$$(1/6)^{18}(5/6)^{82}, \tag{4.1}$$

using independence. Similarly, the probability of rolling 2 non-sixes, then 9 sixes, then 14 non-sixes, then 9 more sixes, and finally 66 non-sixes also has probability given by (4.1). In fact, for every choice of 18 positions in which the sixes may fall, there is a an

outcome with exactly 18 sixes whose probability is $(1/6)^{18}(5/6)^{82}$. Since these outcomes are disjoint, and since there are $\binom{100}{18}$ of them the probability that one of them occurs is

$$\binom{100}{18}(1/6)^{18}(5/6)^{82}. \qquad \square$$

More generally, n independent trials, each having probability p of success and probability $1-p$ of failure, will lead to k total successes with probability

$$\binom{n}{k}p^k(1-p)^{n-k}.$$

We refer to such an experiment as an experiment of *Bernoulli trials*, and the quantities $\binom{n}{k}p^k(1-p)^{n-k}$ are called *binomial probabilities*.

4.2 Hypothesis testing

We can use binomial probabilities to answer questions about experiments in medicine or social science. For example, suppose that currently available drugs successfully treat a disease with probability $p = 0.4$. Suppose that a trial is conducted for an experimental drug, and 48 out of the 100 patients responded positively to the drug. That sounds like an improvement, but perhaps it is due to random chance rather than an actual improvement in the drug.

To assess such situations, researchers use a framework called *hypothesis testing*. One posits a *null hypothesis*, which in this case is that the drug is 40% successful, like currently available drugs. The *alternative* hypothesis is that the drug is more than 40% successful. We answer the question "How likely would we have been to see data at least as extreme as what we observed, if we assume the null hypothesis?" If the answer is greater than some threshold, typically 5%, then we do not have enough evidence to reject the null hypothesis. Otherwise, the experiment is viewed as providing evidence to reject the null hypothesis.

In this case, the probability that 48 or more of 100 subjects would respond positively to the drug is equal to

$$\sum_{k=48}^{100}\binom{100}{k}(0.4)^k(0.6)^{n-k}$$

under the null hypothesis. We can compute this quantity using some lines of code like

```
n = big(100)
m = 48
p = 0.4
sum([binomial(n,k) * p^k * (1-p)^(n-k) for k=m:n])
```

in Julia (we use the function `big` because `binomial` causes an overflow if we try to use default 64-bit integers). It works out to approximately 0.0637, so we don't have quite enough evidence to reject the null hypothesis.

4.3 Inclusion-exclusion and the hat check problem

Suppose (Ω, P) is a discrete probability space. If events E and F are disjoint, then by additivity, we have $P(E \cup F) = P(E) + P(F)$. If E and F are not disjoint, then $P(E) + P(F)$ may be larger than $P(E \cup F)$, because the mass associated to any ω's which are in E and F is added twice in the computation of $P(E) + P(F)$, while they're only added once for $P(E \cup F)$. To relate $P(E \cup F)$ to $P(E)$ and $P(F)$, we have to subtract off these values so that they're only counted once:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

This is called the *principle of inclusion-exclusion* for two events.

Exercise 4.2. Show that the principle of inclusion-exclusion for two events holds in an arbitrary probability space (Ω, P). In other words, use the basic properties of a probability measure to derive the equality $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

For three events A, B, and C, there are more steps (the reader is encouraged to draw a Venn diagram). When we calculate $P(A) + P(B) + P(C)$, the probability mass associated to elements which are in at least two of the sets have been added twice. However, if we subtract the probabilities of all the pairwise intersections, then elements in $A \cap B \cap C$ have been added three times and then subtracted three times. So we have to add $P(A \cap B \cap C)$, giving

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

The inclusion-exclusion principle for n events works similarly: we add all the probabilities of the events, subtract the sum of the probabilities of all the pairwise intersections, then add back the sum of the probabilities of all the three-wise intersections, continuing in this way until we reach the single n-wise intersection of all n sets. See page 104 in your book for a proof of this fact in the discrete case using the binomial theorem. The result may also be proved in the general case by a brute force induction on n. We will see a slick one-line proof later when we discuss expectation.

The inclusion-exclusion formula can be a bit unwieldy, and in the abscence of some nice symmetries, it will usually be unuseful. However, sometimes it works out very nicely:

Example 4.3. A total of n people arrive at a party and check their hats at the door. When they leave, each person collects a hat uniformly at random from the hats not yet collected. What is the probability that no one ends up with the hat that they brought?

Solution. We define Ω to be the set of permutations of $\{1, 2, \ldots, n\}$, where the index in position k of a permutation $\omega \in \Omega$ is to be interpreted as the index of the person whose hat ended up in the possession of person k. Note that the probability measure described in the problem statemen is the uniform measure on Ω. We define for each k between 1 and n the event $E_k \subset \Omega$ that person k receives their own hat; in terms of permutations,

$$E_k = \{\sigma \in \Omega : \sigma(k) = k\}.$$

The probability that someone gets their own hat is the probability of $\bigcup_{k=1}^{n} E_k$, which we may rewrite in terms of probabilities of j-wise intersections of the E_k's using the principle of inclusion-exclusion. These may be evaluated by observing that for all $1 \leq k_1 < \cdots < k_j \leq n$, we have that $P(E_{k_1} \cap E_{k_2} \cap \cdots \cap E_{k_j})$ is equal to $1/(n)_j$ (exercise). Since there are $\binom{n}{j} = (n)_j/j!$ total j-wise intersections, we get

$$P\left(\bigcup_{k=1}^{n} A_k\right) = 1 - 1/2! + 1/3! - \cdots + (-1)^{n+1}/k!.$$

The probability that no one gets their own hat is 1 minus this probability, which works out to

$$\frac{1}{2!} - \frac{1}{3!} + \frac{1}{3!} - \cdots + (-1)^k/k!,$$

which we recognize as a truncation of the Taylor Series for the exponential function, evaluated at -1. Therefore, the desired probability converges to e^{-1} as $k \to \infty$. In fact, it's so close that for all $n \geq 1$, the number of *derangements* (that is, permutations σ with no value of k satisfying $\sigma(k) = k$) is equal to $n!/e$ rounded to the nearest integer. (!) □

4.4 Conditional probabilities in criminal justice: the Sally Clark case

Consider the following story, copied from Wikipedia's article *Sally Clark*:

Sally Clark (August 1964 – 15 March 2007) was a British solicitor who, in November 1999, became the victim of a miscarriage of justice when she was found guilty of the murder of two of her sons. Although the conviction was overturned and she was freed from prison in 2003, she developed serious psychiatric problems and died in her home in March 2007 from alcohol poisoning.

Clark's first son died suddenly within a few weeks of his birth in September 1996, and in December 1998 her second died in a similar manner. A month later, she was arrested and subsequently tried for the murder of both children. The prosecution case relied on statistical evidence presented by paediatrician Professor Sir Roy Meadow, who testified that the chance of two children from an affluent family suffering sudden infant death syndrome was 1 in 73 million. He had arrived at this figure by squaring 1 in 8500, as being the likelihood of a cot death in similar circumstances.

There are at least three statistical criticisms that can be levied against this 1 in 73 million figure.

1. Squaring 1/8500 is based on an assumption of independence of the deaths of the two children. There is not enough known medically about SIDS to infer independence, and in the absence of evidence to the contrary it doesn't seem that reasonable.
2. The event purported to have probability 1/73,000,000 is the event that Sally Clark had two children die from SIDS. However, since there are many British mothers, any one of whom would have presumably been put on trial under the same circumstances, a more relevant probability would be that of the event that *there exists* a British mother who has suffered such a tragedy. This would make the probability larger by a factor approximately equal to the number of British mothers, which is on the order of millions.

   To see why this is important, consider a criminal justice policy that permits a prosecuting case based upon (1) something very unlikely that happened to the defendant, and (2) an explanation of that unlikely event that, if true, would result in a much larger probability of that event. Substitute winning the lottery and cheating the lottery for (1) and (2), and you'll see that such a policy is unreasonable. If, however, we replace "very unlikely" with "so unlikely that it would almost certainly not happen to anyone on earth in the next thousand years", then this line of reasoning does become tenable. Presumbly lottery organizers do apply this kind of analysis to identify participants who may be cheating.
3. Finally, the probability calculated by Professor Meadow is the probability of the evidence (namely, two infant deaths, denoted E) given innocence (I). The quantity more relevant for a trial is the probability of *innocence given the evidence*. These quantities are not the same:

$$P(I|E) = \frac{P(I \cap E)}{P(E)} = \frac{P(I \cap E)}{P(I)} \cdot \frac{P(I)}{P(E)} = P(E|I)\frac{P(I)}{P(E)}.$$

   Since $P(I)$ is close to 1[11], $P(I|E)$ is larger than $P(E|I)$ by a factor of approximately $P(E)$. This quantity includes deaths from both SIDS and murder and has not been estimated in the expert analysis. Since $P(E)$ is small, any conflation of $P(I|E)$ with $P(E|I)$ is extremely misleading.

[11] $P(I)$ denotes the unconditional probability of innocence, that is, the probability that someone is innocent if you don't know anything about their children.

5 Important probability distributions

5.1 Discrete distributions

We describe a few important distributions and discuss how to sample from them using a computer.

5.1.1 Uniform discrete distribution

To generate a random element of $\{1, 2, \ldots, n\}$, we can use the code

```
ceil(n*rand())
```

where `rand()` returns a uniform random number between 0 and 1 and `ceil` is the ceiling function, which returns the least integer which is greater than or equal to its argument. We denote the uniform distribution on a set S as $\text{Unif}(S)$.

5.1.2 Binomial distribution

To generate a random element of $\{0, 1, \ldots, n\}$ from the probability mass function

$$\binom{n}{k} p^k (1-p)^{n-k},$$

we can sum n independent random variables each of which is 1 with probability p and 0 with probability $1 - p$. We will denote this distribution by $\text{Bin}(n, p)$.

5.1.3 Geometric distribution

Consider flipping a weighted coin with probability p of turning up heads, until the flip T when the first head turns up. So if we got TTH, then we would say $T = 3$. Then we have

$$P(T = 1) = p, \quad P(T = 2) = p(1-p), \quad P(T = 3) = p(1-p)^2, \ldots$$

Exercise 5.1. Show that the function whose value at $k \in \{1, 2, 3, \ldots\}$ is $p(1-p)^{k-1}$ does indeed define a probability mass function.

Clearly, it is possible to sample from the geometric distribution by running a sequence of Bernoulli trials with probability p and stopping at the first success. However, if p is very small, then this may take a long time. The following exercise gives a faster approach.

Exercise 5.2. Show that if U is a uniform random variable in $[0,1]$ and Y is the least integer satisfying $1-(1-p)^Y \geq U$, then Y is geometrically distributed. Conclude that

```
ceil(log(rand())/log(1-p))
```

returns a random variable which is geometrically distributed.

We denote the geometric distribution by $\text{Geom}(p)$.

5.1.4 Negative binomial distribution

Rather than waiting until we see the *first* success in a sequence of Bernoulli trials, let's continue until the time T when we see the kth success, where $k \geq 1$ is an integer. Then T is said to have *negative binomial distribution*. We can think of waiting for the kth success as k independent repetitions of the experiment where we wait for the first success. This leads to the following fast way algorithm for sampling from the negative binomial distribution.

Exercise 5.3. Show that if $U_1, \ldots, U_k$ are independent uniform random variables in $[0,1]$ and Y_i is the least integer satisfying $1-(1-p)^{Y_i} \geq U$, then $\sum_{i=1}^k Y_i$ has negative binomial distribution.

Exercise 5.4. Use a counting argument to show that if X is has negative binomial distribution with parameters p and k, then

$$P(X=n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}.$$

5.1.5 Poisson distribution

Consider a Bernoulli trials experiment with a very large number n of trials, but with a very small success probability $p(n)$, chosen so that the number of successes goes neither to 0 nor to ∞ as $n \to \infty$. More precisely, we take $p(n) = \lambda/n$, where $\lambda > 0$ is a constant. Then if X is the number of successes, then for small values of k,

$$\begin{aligned} P(X=k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &\sim \frac{n^k}{k!} \frac{\lambda^k}{n^k} e^{-\lambda} \\ &= \boxed{\frac{\lambda^k}{k!} e^{-\lambda}}. \end{aligned}$$

We can use this expression to define a probability measure on the set of nonnegative integers:

Exercise 5.5. Show that $m(k) = \frac{\lambda^k}{k!}e^{-\lambda}$ defines a probability mass function on $\{0, 1, 2, \ldots\}$.

The probability measure associated with m is called the *Poisson distribution*, denoted $\text{Poiss}(\lambda)$. We may summarize the relationship between the Poisson and binomial distributions by saying that if n is large and p is small, then $\text{Bin}(n, p)$ is well approximated for small values of k by the Poisson distribution with parameter $\lambda = np$.

Example 5.6. Suppose that there are 100 total points sprinkled independently and uniformly in the square $[0, 5] \times [0, 5]$. Consider a region $R \subset [0, 5] \times [0, 5]$ whose area is 0.1. Use the Poisson distribution to estimate the probability that exactly 2 of the 100 points fall in the region R.

Solution. The number of points lying in the region R has the binomial distribution with parameters $n = 100$ and $p = 0.1/25 = 0.004$. Thus the probability that exactly two points fall in the region R is given by

$$\binom{100}{2}(0.004)^2(0.996)^{98} \approx 0.05347$$

Using the Poisson distribution, we note that $p = \lambda/n$ for $\lambda = 0.4$, so we estimate the probability mass at $k = 2$ to be

$$\frac{\lambda^2}{2!}e^{-\lambda} \approx 0.05363.$$

This approximation is quite accurate. □

5.1.6 Benford's law

Suppose we take a large set of real-world data spanning several orders of magnitude and tally the resulting initial digits. One might expect to see a histogram matching the uniform distribution on $\{1, 2, \ldots, 9\}$, but in practice 1 is by far the most common initial digit, and the relative frequencies decrease from 1 to 9 (see Figure 1).

The model most often used for describing these relative frequencies is called *Benford's distribution*, defined by assigning probability

$$\log_{10}(k + 1) - \log_{10}(k)$$

to each integer value of k from 1 to 9. The phenomenon of real-world initial digits data closely matching Benford's distribution is called *Benford's law*.

According to Wikipedia, Benford's law does not have a thoroughly convincing mathematical explanation. In some sense it cannot, since the "real world data" part of Benford's law admits no rigorous mathematical formulation. However, Exercise 5.7 shows how Benford's distribution emerges if the logarithm of a random variable is uniformly

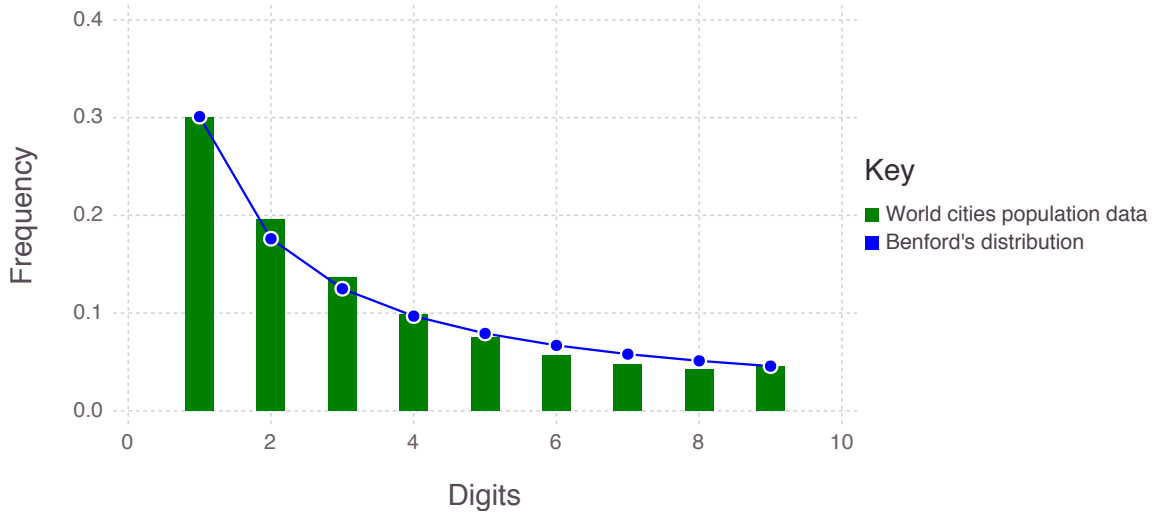

Figure 1: A histogram of initial digits of the populations of 47,890 cities from around the world, compared to the prediction made by Benford's law.

distributed. This is a reasonable assumption for any quantity which grows exponentially (such as population), although Benford's law sometimes works for data with no obvious connection to exponential growth (like entries in a large company's accounting data).

Exercise 5.7. Show that if U is uniform on $[1,2]$, then the law of the tens digit (that is, the first digit) of 10^U is Benford's distribution.

5.2 Continuous distributions

In this section we discuss some common probability density functions and their properties. Recall from Exercise 1.8 that if we have $Y = g(X)$ where g is an increasing function and X and Y are random variables with pdfs f_X and f_Y, then the cdfs F_X and F_Y of X and Y (respectively) are related by

$$F_Y(y) = F_X(g^{-1}(y)). \tag{5.1}$$

Taking derivatives, we find that

$$f_Y(y) = f_X(g^{-1}(y))(g^{-1})'(y). \tag{5.2}$$

The following observation is very useful for simulating random variables with arbitrary cdfs.

Exercise 5.8. Show that if $U \sim \text{Unif}([0,1])$ and $F : \mathbb{R} \to [0,1]$ is an increasing[12], right-continuous function, then the cdf of $F^{-1}(U)$ is equal to F.

5.2.1 Uniform density on $[a,b]$

The uniform density on $[a,b]$ is defined to be the constant function $f(x) = 1/(b-a)$ for all $x \in [a,b]$, and is denoted $\text{Unif}([a,b])$. We can sample from this distribution with following code:

```
a + (b-a)*rand()
```

5.2.2 Exponential density

The function $f(x) = \lambda e^{-\lambda x}$ defined on $[0,\infty)$ is called the exponential density, denoted $\text{Exp}(\lambda)$. As we saw in Exercise 3.1, the exponential density has the following memorylessness property: if $T \sim \text{Exp}(\lambda)$, then

$$P(T > t+s | T > s) = P(T > t).$$

By Exercise 5.8, the following code returns an exponentially distributed random number with parameter `lambda`:

```
-log(rand())/lambda
```

The exponential density and the Poisson distribution are related in the following way: if we let $(X_i)_{i=1}^{\infty}$ to be an independent sequence of random variables with $X_i \sim \text{Exp}(\lambda)$ and define $S_n = X_1 + \ldots + X_n$, then the number of values of n such that $S_n \leq t$ is Poisson distributed with parameter λt.

5.2.3 Gaussian density

For $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we define the Gaussian distribution, denoted $\mathcal{N}(\mu,\sigma)$, to be the probability measure on $\mathbb{R}$ whose density function is

$$\boxed{f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.}$$

See Figure 2 for a plot. To show that f indeed integrates to 1, one can write $\left(\int_{-\infty}^{\infty} f(x)\,dx\right)^2$ as $\iint_{\mathbb{R}^2} f(x)f(y)\,dx\,dy$ and switch to polar coordinates (exercise). The Gaussian distribution is also called the *normal* distribution, and a normal distribution with $\mu = 0$ and $\sigma = 1$ is called *standard*.

[12] This is true even if F is only weakly increasing, assuming one interprets F^{-1} as the *generalized inverse* of F, defined by $F^{-1}(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}$.

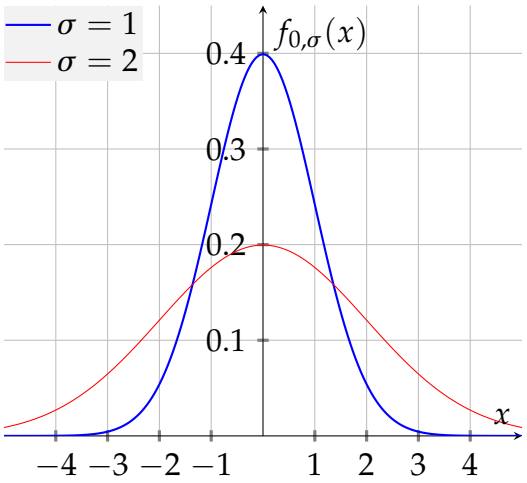

Figure 2: The Gaussian density function $f_{\mu,\sigma}$ with $\mu = 0$ and two different values of σ.

Loosely speaking, the parameter μ specifies the center of the density function (the point about which $f_{\mu,\sigma}$ has a reflection symmetry), and the parameter σ dictates how spread out the density function is. We will identify μ and σ more precisely when we develop the concepts of mean and variance.

The following exercise shows that to generate random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, it suffices to generate random variables with distribution $\mathcal{N}(0, 1)$.

To generate normal random variables, we use the fact (whose proof we omit) that if U and V are independent and uniformly distributed in $[0, 1]$, then $\sqrt{\log(1/U^2)}\cos(2\pi V)$ and $\sqrt{\log(1/U^2)}\sin(2\pi V)$ are independent $\mathcal{N}(0, 1)$ random variables.

Exercise 5.9. Show that if $Z \sim \mathcal{N}(0, 1)$, then $\sigma Z + \mu \sim \mathcal{N}(\mu, \sigma)$.

5.2.4 Cauchy density

Consider directing a ray from the point $(0, -1)$ at an angle Θ selected uniformly at random from $[0, \pi]$, measured with respect to the vector $\langle 1, 0\rangle$. Define X to be the x-coordinate of the point where this ray intersects the x-axis.

Exercise 5.10. Show that the law of X has density

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

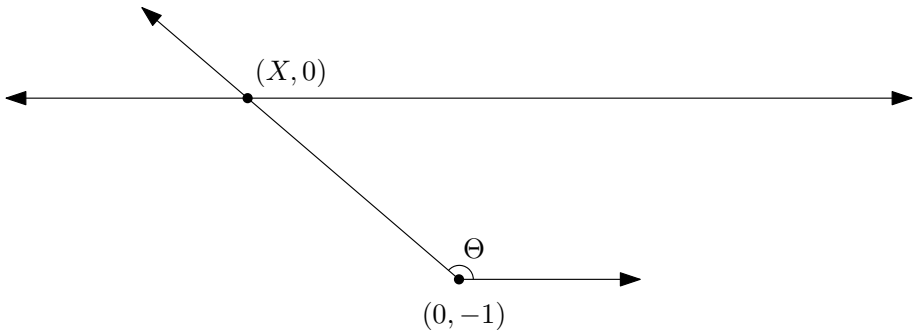

Figure 3: An illustration of how to define a Cauchy random variable in terms of a uniform random variable Θ.

6 Expectation and variance of discrete random variables

6.1 Introduction

Let (Ω, P) be a probability space. The *expectation* of a random variable $X : \Omega \to R$, where $R \subset \mathbb{R}$ is countable, is defined to be a weighted average of the numbers in R, with weights given by the law of X. More precisely, the expectation $E(X)$ is defined to be

$$\boxed{E(X) = \sum_{x \in R} x\, P(X = x).} \tag{6.1}$$

We will see that (under some technical conditions) the expectation of a random variable coincides with another way of averaging: loosely speaking, the average of many independent samples from the law of X is approximately equal to $E(X)$ with high probability.

Example 6.1. Let X be the number of heads in two independent, fair coin flips. Then the expectation of X is given by

$$E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) = 0 + 1/2 + 2 \cdot 1/4 = 1.$$

Example 6.2. Let $X \sim \text{Geom}(1/2)$. Then $E(X) = \sum_{n=1}^{\infty} n\, P(X = n) = \sum_{n=1}^{\infty} n 2^{-n}$. To evaluate this infinite sum, we notice that we can write it as

$$\begin{aligned} &1/2 + 1/4 + 1/8 + 1/6 + \cdots \\ &\quad + 1/4 + 1/8 + 1/6 + \cdots \\ &\quad\quad + 1/8 + 1/6 + \cdots, \end{aligned}$$

since adding these terms up in columns gives the desired sum. The sum of the kth row[13] is equal to 2^{k-1}, so the sum of all the rows is $\sum_{k=1}^{\infty} 2^{k-1} = 2$.

[13]Switching from columns to rows does not change the result because of the following theorem from analysis: if $a_{i,j} \geq 0$ for all i, j then $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} a_{i,j} = \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} a_{i,j}$.

Example 6.3. Let $Y = 2^X$, where $X \sim \text{Geom}(1/2)$. Then $E(Y)$ is given by the sum $\sum_k kP(Y = k)$, where k ranges over the possible values of Y, which are the integer powers of 2. By rewriting $k = 2^n$, this sum is equal to

$$E(Y) = \sum_{n=1}^{\infty} 2^n P(2^X = 2^n) = \sum_{n=1}^{\infty} 2^n / 2^n = \infty.$$

Notice in the previous problem that, for the function $\phi(x) = 2^x$, we have

$$E(\phi(X)) = \sum_{x \in R} \phi(x)\, P(X = x). \tag{6.2}$$

In other words, to find the expectation of $\phi(X)$, we replace x with $\phi(x)$ in (6.1). This is true in general; we regroup terms in the definition of $E(\phi(X))$ as follows:

$$\begin{aligned} E(\phi(X)) &= \sum_{y \in \phi(R)} y\, P(\phi(X) = y) \\ &= \sum_{y \in \phi(R)} y \sum_{x\,:\,\phi(x)=y} P(X = x) \\ &= \sum_{y \in \phi(R)} \sum_{x\,:\,\phi(x)=y} y\, P(X = x) \\ &= \sum_{y \in \phi(R)} \sum_{x\,:\,\phi(x)=y} \phi(x)\, P(X = x) \\ &= \sum_{x \in R} \phi(x)\, P(X = x). \end{aligned}$$

6.2 Linearity of expectation

Let X be the number of fixed points[14] of a random permutation on $\{1, 2, \ldots, n\}$. Calculating $E(X)$ using the definition of expectation would be difficult because it would require determining, for each value of $1 \leq k \leq n$, the probability that the number of fixed points is equal to k. However, we can write

$$X = X_1 + X_2 + \cdots + X_n\,,$$

where X_k is equal to 1 if k is a fixed point and 0 otherwise.

Exercise 6.4. Show that $E(X) = E(X_1) + \cdots + E(X_n)$ by direct calculation, in the cases $n = 2$ and $n = 3$.

Exercise 6.4 suggests that perhaps E distributes across addition. This is indeed true, and it is perhaps more surprising than the notation might suggest, because the distribution

[14] Recall that a fixed point of a permutation σ is an integer k such that $\sigma(k) = k$.

of $X+Y$ depends[15] on the joint distribution of $X+Y$, whereas $E(X)+E(Y)$ can be calculated knowing only the distributions of X and Y individually (in other words, the marginal distributions of X and Y).

Theorem 6.5. Let $c \in \mathbb{R}$. For all discrete random variables X and Y, we have

$$E(X+Y) = E(X) + E(Y) \quad \text{and} \quad E(cX) = cE(X).$$

Proof. Let R_1 be the range[16] of X and let R_2 be the range of Y. Using (6.2), we have

$$\begin{aligned} E(X+Y) &= \sum_{z \in \mathbb{R}} z\, P(X+Y=z) \\ &= \sum_{z \in \mathbb{R}} \sum_{\substack{x \in R_1 \\ y \in R_2 \\ \text{such that } x+y=z}} z\, P(X=x \text{ and } Y=y), \end{aligned}$$

because the event $\{X+Y=z\}$ is a disjoint union of the events $\{X=x\} \cap \{Y=y\}$ where (x,y) ranges over all pairs such that $x+y=z$. Writing z as $x+y$, we can rewrite the double sum as a single sum over all pairs (x,y) with $x \in R_1$ and $y \in R_2$. We get

$$\begin{aligned} &= \sum_{x \in R_1, y \in R_2} (x+y)\, P(X=x \text{ and } Y=y) \\ &= \sum_{x \in R_1, y \in R_2} x\, P(X=x \text{ and } Y=y) + \sum_{x \in R_1, y \in R_2} y\, P(X=x \text{ and } Y=y) \\ &= \sum_{x \in R_1} x\, P(X=x) + \sum_{y \in R_2} y\, P(Y=y). \end{aligned}$$

where the last step follows from the fact that $P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$ whenever A and $(B_i)_{i=1}^{\infty}$ are events and the sequence $(B_i)_{i=1}^{\infty}$ is pairwise disjoint.

The proof of $E(cX) = cE(X)$ is similar but more straightforward. □

Example 6.6. Shuffle a standard 52-card deck, and let X be the number of consecutive pairs of cards in the deck which are both red. Find $E(X)$.

Solution. Calculating $E(X)$ from the distribution of X would be a big mess, because the distribution of X is complicated. Try to calculate, for instance, $P(X=10)$.

However, finding the mean is nevertheless manageable: we can write $X = X_2 + \ldots + X_{52}$ where X_j is equal to 1 if the cards in positions $j-1$ and j are both red and is equal to 0 otherwise[17]. We see that $E(X_j) = (1/2)(25/51)$, since card j is red with probability

[15] For concreteness, consider on one hand two fair independent dice rolls X and Y, and on the other hand a die roll X and a random variable Y defined to be the same as X.

[16] That is, $R_1 = \{r \in \mathbb{R} : P(X=r) > 0\}$.

[17] The X_j's are examples of what we call *indicator* random variables, which are random variables taking values in $\{0,1\}$. They are in one-to-one correspondence with events $E \subset \Omega$ (exercise).

$1/2$, and card $j-1$ is red with conditional probability $25/51$, given that card j is red. So $E(X) = E(X_2) + \cdots + E(X_{52}) = 51(1/2)(25/51) = 25/2$. □

Example 6.7. Let X be the number of records in a random permutation σ on $\{1, 2, \ldots, n\}$, that is, the number of integers k such that $\sigma(k) > \sigma(j)$ for all $j < k$. Then

$$X = X_1 + \cdots + X_n$$

where X_k is equal to 1 if k is a record and is equal to 0 otherwise. Thus $E(X) = \sum_{k=1}^{n} E(X_k)$. The expectation of X_k is equal to the probability that k is a record, by the definition of expectation. The largest element in the list $\sigma(1), \ldots, \sigma(k)$ is equally likely to be in any of the k positions (exercise). Therefore, $E(X_k) = 1/k$, and we get

$$E(X) = 1 + 1/2 + \cdots + 1/n,$$

which is asymptotic to $\log n$ as $n \to \infty$, since it's a left-endpoint Riemann sum for the integral of $1/x$ from 1 to ∞.

Exercise 6.8. If X and Y are discrete, independent random variables, then $E(XY) = E(X)E(Y)$.

6.3 Conditional expectation

The conditional expectation $E(X \mid F)$ of a discrete random variable X given an event F is defined in the same way as the expectation of X, except the probability measure P is replaced by the conditional probability given F:

$$E(X \mid F) = \sum_{x \in \mathbb{R}} x\, P(X = x \mid F).$$

Example 6.9. Let X be a random variable whose law is given by the uniform measure on $\{1, 2, 3, 4, 5, 6\}$. Find the conditional expectation of X given the event $X \neq 5$.

Solution. By the definition of conditional probability, $P(X = k \mid X \neq 5) = 1/5$ for all $k \in \{1, 2, 3, 4, 6\}$. Thus

$$E(X \mid X \neq 5) = 1/5 + 2/5 + 3/5 + 4/5 + 6/5 = 16/5.$$ □

We can use the idea of conditional expectation to formalize the concept of a fair game. For example, consider a game where a fair coin is flipped repeatedly, and the player wins \$1 for every head and loses \$1 for every tail. This game is fair, because the expected amount of money we win on the next step is zero, regardless of the results of the preceding coin flips. In other words, if we denote by X_k the amount of money won after the kth flip, then we have

$$E(X_{n+1} - X_n \mid \{X_1 = x_1, \ldots, X_n = x_n\}) = 0. \tag{6.3}$$

for all n and for all $x_1, x_2, \ldots, x_n$ such that the event $\{X_1 = x_1, \ldots, X_n = x_n\}$ has positive probability. More generally, a *martingale* is a sequence of random variables $X_1, X_2, \ldots$ that satisfy (6.3).

6.4 Variance

The variance of a random variable X measures the extent to which the distribution of X is concentrated (in which case the variance is small) or spread out (in which case the variance is large). For example, a random variable which equals 100.1 with probability $1/2$ and 99.9 with probability $1/2$ has a smaller variance than a random variable which equals 10 with probability $1/2$ and -10 with probability $1/2$.

Although there are other reasonable ways to measure how spread out a distribution is, the one with the nicest properties is the average squared difference between X and the mean of X. Thus, we define the variance $\mathrm{Var}(X)$ of X to be

$$\boxed{\mathrm{Var}(X) = E((X - E(X))^2) = E((X - \mu)^2)},$$

where μ is an abbreviation for $E(X)$.

Proposition 6.10. $\mathrm{Var}(X) = E(X^2) - E(X)^2$

Proof. Squaring and applying linearity of expectation, we have

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2E(X)\mu + \mu^2 \\ &= E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2,\end{aligned}$$

as desired. □

Exercise 6.11. Let X be a discrete random variable. Show that for all $c \in \mathbb{R}$, we have $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$, and for all **independent** discrete random variables X and Y, we have $\mathrm{Var}(X) + \mathrm{Var}(Y)$.

If the random variable X is interpreted as a physical quantity associated with some unit, then the variance must be associated with the square of that unit (for example, if X represents a number of meters, the $\mathrm{Var}(X)$ represents a number of square meters). To define a quantity which encodes the variance of X but which is commensurable with X, we introduce the standard deviation

$$\sigma(X) = \sqrt{\mathrm{Var}(X)}.$$

Proposition 6.12. Let $X_1, X_2, \ldots, X_n$ be independent random variables which all have the same law[18]. Denote by σ and μ the standard deviation and mean of X_1, and assume σ and μ exist and are finite. Define random variable $A_n = (X_1 + \ldots + X_n)/n$. Then

$$\begin{aligned}E(A_n) &= \mu \\ \sigma(A_n) &= \frac{\sigma}{\sqrt{n}}.\end{aligned}$$

[18]Hereafter, and throughout the literature, abbreviated as *i.i.d.*, for *independent and identically distributed*.

Exercise 6.13. Prove Proposition 6.12.

Proposition 6.12 carries an important philosophical interpretation: if we average successive samples from a distribution whose mean we wish to estimate, we are indeed estimating the mean (since $E(A_n) = \mu$), and we are doing so with increasing accuracy (since $\sigma(A_n) \to 0$ as $n \to \infty$).

7 Expectation and variance of continuous random variables

7.1 Expectation of an arbitrary random variable

Many ideas from discrete probability transfer to continuous probability via a natural limiting procedure. The general idea is to approximate a continuous random variable with a discrete random variable in a way that becomes increasingly accurate as some parameter n goes to ∞.

For example, suppose we want to find the expected value of random variable X with distribution $\text{Exp}(1)$. Our formula $\sum_{x\in\mathbb{R}} x\,P(X = x)$ does not apply since the summand is always equal to 0. However, we may define for each integer $n \geq 0$ the random variable which maps each $\omega \in \Omega$ to

$$X_n(\omega) = 2^{-n}\lfloor 2^n X(\omega)\rfloor.$$

Note that the right hand side is the same as "$X(\omega)$ rounded down to the nearest integer multiple of 2^{-n}." So, for example, if $X(\omega) = \pi$, then $X_0(\omega) = 3$ and $X_3(\omega) = 3\frac{1}{8}$. Clearly, with this definition, X_n and X never differ by more than 2^{-n}. Furthermore, since X_n is a discrete random variable, we already know how to define its expectation. Thus it is reasonable to define

$$E(X) = \lim_{n\to\infty} E(X_n). \tag{7.1}$$

Note that X_n is an increasing sequence of random variables (meaning that $X_n(\omega) \leq X_{n+1}(\omega)$ for all $\omega \in \Omega$ and $n \geq 0$), which implies that $E(X_n)$ is an increasing sequence of real numbers. Every increasing sequence of numbers either converges to a finite limit or diverges to $+\infty$. Thus, if we admit $+\infty$ as a permissable limiting value, $E(X)$ exists and equals some number in the interval $[0, +\infty]$.

There are couple of aspects of (7.1) which are important and unsurprising: (1) the limit is the same even if we replace 2^{-n} with some other sequence that tends to 0, and (2) if X is already a discrete random variable, then the limit on the right hand side is equal to $E(X)$ as calculated from our original definition of expectation. Thus we are justified in using the symbol E for both notions of expectation. In fact, (7.1) works for *any* nonnegative

random variable, not just discrete or continuous ones[19].

For random variables X which are not necessarily nonnegative, we define the positive and negative parts

$$X^+(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0 \\ 0 & \text{if } X(\omega) < 0, \end{cases}$$

and

$$X^-(\omega) = \begin{cases} -X(\omega) & \text{if } X(\omega) < 0 \\ 0 & \text{if } X(\omega) < 0. \end{cases}$$

These are both nonnegative random variables, and $X = X^+ - X^-$ (exercise). Thus it is reasonable to define

$$E(X) = E(X^+) - E(X^-).$$

as long as $E(X^+)$ and $E(X^-)$ are not both equal to $+\infty$. In that case, we say that the expectation of X is not defined.

Proposition 7.1. If X and Y are random variables with finite mean, then $E(X+Y) = E(X) + E(Y)$ and $E(cX) = cE(X)$ for all real numbers $c \in \mathbb{R}$.

Proof idea. Apply the discrete version of linearity of expectation to approximating sequences X_n and Y_n which converge to X and Y, and take $n \to \infty$. □

7.2 Expectation of a continuous random variable

Suppose that X is a random variable with pdf f, and suppose that $f(x) = 0$ for all $x \leq 0$. Denote by D_n the set of nonnegative integer multiples of 2^{-n}, and note that

$$E(X) = \lim_{n\to\infty} E(X_n) = \lim_{n\to\infty} \sum_{x\in D_n} x\, P(X_n = x),$$

where X_n is X rounded down to the nearest element of D_n. Note that the event $\{X_n = x\}$ is the same as the event $\{X \in [x, x+2^{-n})\}$. To find this probability, we integrate f over $[x, x+2^{-n})$, and our expression for the expectation becomes

$$\lim_{n\to\infty} \sum_{x\in D_n} x \int_x^{x+2^{-n}} f(t)\, dt. = \lim_{n\to\infty} \sum_{x\in D_n} \int_x^{x+2^{-n}} tf(t)\, dt,$$

where we have used the fact that $t \approx x$ for all t between x and $x+2^{-n}$, with an error that tends to 0 as $n \to \infty$[20]. This sum now telescopes, and we're left with

$$E(X) = \int_0^\infty xf(x)\, dx.$$

[19] As an example of a random variable Z which is neither discrete nor continuous, let X be a die roll and Y an independent normal random variable, and let $Z(\omega)$ equal $X(\omega)$ if $X(\omega)$ is 1, 2, or 3, and set $Z(\omega) = Y(\omega)$ otherwise.

[20] We're playing a little fast and loose here.

If f is positive somewhere on the negative real line, then we get

$$E(X^+) = \int_0^\infty xf(x)\,dx \text{ and } E(X^-) = \int_{-\infty}^0 xf(x)\,dx,$$

which imply

$$\boxed{E(X) = \int_{-\infty}^\infty xf(x)\,dx,}$$

and this is our formula for the expectation of a random variable X with density f.

Example 7.2. Show that the expectation of a Unif$([0,1])$ random variable is $1/2$.

Solution. The density of a Unif$([0,1])$ random variable X is equal to[21] $\mathbf{1}_{[0,1]}$, so

$$E(X) = \int_{-\infty}^\infty x\mathbf{1}_{[0,1]}(x)\,dx = \int_0^1 x\,dx = 1/2. \qquad \square$$

The following theorem enables us to use the density of X to calculate the expectation of random variables defined in terms of X.

Theorem 7.3. If $\phi : \mathbb{R} \to \mathbb{R}$ is continuous and X has density f, then

$$E(\phi(X)) = \int_{-\infty}^\infty \phi(x)f(x)\,dx. \tag{7.2}$$

Proof idea. For simplicity, assume that X and ϕ are both nonnegative. By definition, we have

$$E(\phi(X)) = \lim_{n\to\infty} E(2^{-n}\lfloor 2^n\phi(X)\rfloor) = \lim_{n\to\infty} E(\phi(2^{-n}\lfloor 2^n X\rfloor)),$$

where the second equality, although it has not been rigorously justified here, makes sense because ϕ is continuous. Now

$$E(\phi(2^{-n}\lfloor 2^n(X)\rfloor)) = \sum_{x\in D_n} \phi(x)\int_x^{x+2^{-n}} f(t)\,dt,$$

by (6.2), which is the discrete version of the present theorem. Writing $\phi(x) \approx \phi(t)$ (again using continuity) and telescoping the sum, we obtain (7.2). $\square$

Proposition 7.4. If X and Y are continuous, independent random variables, then $E(XY) = E(X)E(Y)$.

Exercise 7.5. Use the fact that X and Y are independent if and only if the joint density of X and Y is equal to $f_X(x)f_Y(y)$, where f_X and f_Y are the densities of X and Y, to prove Proposition 7.4.

[21] We denote by $\mathbf{1}_A$ the function which is equal to 1 on the set A and 0 otherwise.

7.3 Variance of a continuous random variable

As in the discrete case, we define variance in terms of expectation via

$$\operatorname{Var} X = E((X - \mu)^2),$$

where $\mu = E(X)$. In light of (7.2), we have

$$\operatorname{Var} X = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx,$$

where f is the density of X.

Note that the properties

1. $\operatorname{Var}(cX) = c^2 \operatorname{Var}(X)$
2. $\operatorname{Var}(X + c) = \operatorname{Var}(X)$
3. $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y)$ if X and Y are independent

all carry over from the discrete case, because we proved them using basic properties of expectation rather than using details specific to the discrete setting.

Exercise 7.6. (a) Show that if $X \sim \operatorname{Unif}([0,1])$, then $\operatorname{Var}(X) = 1/12$.
(b) Show that if $X \sim \operatorname{Exp}(\lambda)$, then $E(X) = 1/\lambda$ and $\operatorname{Var} X = 1/\lambda^2$.
(c) Show that the variance of a $\mathcal{N}(0,1)$ random variable is 1. (Hint: split the factor x^2 into $x \cdot x$ and use integration by parts.)

Exercise 7.7. Show that $\sigma Z + \mu$ has mean μ and variance σ^2, if $Z \sim \mathcal{N}(0,1)$.

Example 7.8. The Cauchy density

$$f(x) = \frac{a}{\pi} \cdot \frac{1}{a^2 + x^2}$$

has undefined mean, because if we multiply by x and try to integrate over $\mathbb{R}$, we get a contribution of $-\infty$ from the negative real line and $+\infty$ from the positive real line. We say that the Cauchy distribution has *heavy tails*, meaning that the probability that $X > x$ goes to 0 slowly as $x \to \infty$ (and similarly for $X < x$ as $x \to -\infty$).

8 Sums of independent random variables

8.1 Sums of integer-valued random variables

Suppose that X and Y are independent random variables with probability mass functions m_1 and m_2 supported[22] on $\mathbb{Z}$. Then

$$P(X+Y=k) = \sum_{j\in\mathbb{Z}} m_1(j)m_2(k-j), \tag{8.1}$$

since the event $\{X+Y=k\}$ is a disjoint union of the events $\{X=j\}\cap\{Y=k-j\}$ as j ranges over $\mathbb{Z}$. Since $X+Y$ is integer-valued, (8.1) fully specifies the law of Z. We use the term *convolution*, denoted $m_1 * m_2$, for the operation on m_1 and m_2 specified in (8.1). So, we can say that the law of the sum of two independent integer-valued random variables is equal to the convolution of their laws.

Example 8.1. Find the convolution of $\mathrm{Unif}(\{1,2,3,4,5,6\})$ with itself.

Solution. Denoting by m the pmf of $\mathrm{Unif}(\{1,2,3,4,5,6\})$, we have $(m*m)(2) = m(1)^2 = 1/36$, $(m*m)(3) = m(1)m(2)+m(2)m(1) = 2/36$, etc, up to $(m*m)(12) = m(6)^2 = 1/36$. □

Example 8.2. Find the convolution of $\mathrm{Bin}(m,p)$ and $\mathrm{Bin}(n,p)$.

Solution. Using the interpretation of $\mathrm{Bin}(m,p)$ as the number of heads in m independent coin flips, we see that adding two independent random variables with laws $\mathrm{Bin}(m,p)$ and $\mathrm{Bin}(n,p)$ gives a random variable with law $\mathrm{Bin}(m+n,p)$. □

8.2 Sums of independent continuous random variables

Suppose that X and Y are independent random variables with densities f and g. Then $X+Y\in(t,t+\Delta t)$ if and only if (X,Y) lies in the thin strip $\{(x,y)\in\mathbb{R}^2 : t < x+y < t+\Delta t\}$. The double integral of the density $f(x)g(y)$ over this strip is given by

$$\int_{-\infty}^{\infty}\int_{t-y}^{t-y+\Delta t} f(x)g(y)\,dx\,dy.$$

Dividing by Δt, taking $\Delta t\to 0$, and applying the fundamental theorem of calculus, we find that the density of $X+Y$ is given by

$$(f*g)(t) = \int_{-\infty}^{\infty} f(t-y)g(y)\,dy,$$

[22] We say that a probability mass function is *supported* on a set $A\subset\mathbb{R}$ if it assigns no mass to the set $\mathbb{R}\setminus A$.

where $f * g$ denotes the convolution of the densities f and g.

Exercise 8.3. Convolve Unif$([0,1])$ with itself.

Exercise 8.4. Convolve Exp(λ) with itself.

As Exercises 8.3 and 8.4 show, the convolution of a density function with itself does not necessarily belong to the same class of distributions (that is, the convolution of two uniform random variables is not uniform, and the convolution of two exponential random variables is not exponential). The Gaussian and Cauchy distributions, however, are closed under convolution:

Proposition 8.5. If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and X and Y are independent, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Remark 8.6. Note that we can show that the mean and variance of $X + Y$ are $\mu_1 + \mu_2$ and $\sigma_1^2 + \sigma_2^2$ using basic properties of expectation and variance. The proposition is interesting because the sum is *Gaussian*.

Proof. This may be done by brute force using the definition of the convolution of two densities and completing the square. However, here's a slicker and more intuitive proof[23].

If we prove the result in the zero-mean case, then we can apply it to the random variables $X - \mu_1$ and $Y - \mu_2$ to obtain the general result. Thus we assume $\mu_1 = \mu_2 = 0$.

Writing $X + Y$ as $\sigma_1 Z_1 + \sigma_2 Z_2$ where Z_1 and Z_2 are standard normal random variables, we see that the cdf of $X + Y$ is

$$F_{X+Y}(t) = \int_{\sigma_1 x + \sigma_2 y \leq t} f(x) f(y)\, dx\, dy,$$

where f is the density of standard normal distribution. Observe that $f(x)f(y)$ is proportional to $\exp(-(x^2 + y^2)/2)$, which in polar coordinates is $\exp(-r^2/2)$. This means that $f(x)f(y)$ is a rotationally symmetric function on $\mathbb{R}^2$. Therefore, we apply a suitable rotation to the domain of integration in the above integral to send the line $\sigma_1 x + \sigma_2 y \leq t$ to a vertical line. Since the distance from the origin to the line $ax + by = c$ is equal to[24] $c\sqrt{a^2 + b^2}$, we get

$$F_{X+Y}(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t\sqrt{\sigma_1^2 + \sigma_2^2}} f(x) f(y)\, dx\, dy,$$

which equals the standard normal cdf evaluated at $t\sqrt{\sigma_1^2 + \sigma_2^2}$. Thus the cdf of $X + Y$ divided by $\sqrt{\sigma_1^2 + \sigma_2^2}$ is the standard normal cdf, and it follows that $X + Y$ is a normal random variable with mean zero and standard deviation $\sqrt{\sigma_1^2 + \sigma_2^2}$. □

[23] http://math.stackexchange.com/questions/228/proof-that-the-sum-of-two-gaussian-variables-is-another-gaussian

[24] This is a geometry exercise with many solutions. One simple one involves calculating the area of the triangle whose vertices are the origin and the two intercepts of the line in two different ways.

Proposition 8.7. If X and Y are Cauchy random variables and $a, b > 0$, then $aX + bY$ has the same law as $a + b$ times a Cauchy random variable.

Proof. The density of aX is $\frac{a}{\pi}\frac{1}{a^2+x^2}$, and similarly for bY. So the law of $aX + bY$ is the convolution

$$f_{aX+bY}(t) = \frac{ab}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{a^2+(t-x)^2}\cdot\frac{1}{b^2+x^2}\,dx.$$

This integral is quite a mess to evaluate analytically, but if we run the SageMath computer algebra system (CAS)[25]

```
var("a b x z") # declares that a,b,x,z are symbols
assume(a>0,b>0) # notifies the integrator that a and b are positive
I = (a*b/pi^2)*integrate(1/((a^2+x^2)*(b^2+(z-x)^2)),
            x,-infinity,infinity)
factor(I) # we factor so the CAS can make appropriate cancellations
```

we find that $f_{aX+bY}(t) = \frac{a+b}{\pi}\frac{1}{(a+b)^2+t^2}$. We obtain the same expression if we calculate the density of $a + b$ times a Cauchy random variable (exercise; use (5.2)), so this concludes the proof. □

Taking $a = b = 1/2$ in Proposition 8.7, we discover the remarkable fact that *the average of two Cauchy random variables is again Cauchy*. In particular, the average of two independent Cauchy random variables isn't any more concentrated around zero than one Cauchy random variable. Perhaps even more surprisingly, no matter how large we choose n, the mean of n independent Cauchy random variables is also a Cauchy random variable with the same distribution!

Contrast this situation with Proposition 6.12, where averages of random variables with finite variance *always* concentrate around the mean. The reason the Cauchy density is not a counterexample to Proposition 6.12 is that the hypothesis that the mean and variance of X exist is not satisfied.

[25] See `https://sagecell.sagemath.org` for a box you can just type code into. The advantage it has over Wolfram | Alpha is that you can run blocks of code, as opposed to just one-liners. Also, W | A happens to choke on this particular calculation. Disadvantage: the system throws an error unless you use correct code. W | A tries to figure out what you meant.

9 Law of large numbers

9.1 Chebyshev's inequality

One's intuition about variance suggests that if the variance of a random variable is small, then that random variable must be pretty likely to be pretty close to its mean. After all, we introduced variance as a way to quantify the idea of a random variable having a large probability of being far from its mean. Chebyshev's inequality is a way of quantifying this intuition.

Theorem 9.1. (Chebyshev's inequality) Let X be a random variable with finite mean and variance. Then for all $\lambda > 0$, we have

$$P(|X - \mu| \geq \lambda) \leq \frac{\operatorname{Var} X}{\lambda^2}. \tag{9.1}$$

Remark 9.2. Replacing λ with $k\sigma$, where σ is the standard deviation of X, we get

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Thus, Chebyshev's inequality says that the probability that a random variable is more than k standard deviations away from its mean is never greater than $1/k^2$.

Before proving Chebyshev's inequality, we point out that for any probability space (Ω, P) and event A, we have

$$P(A) = E(\mathbf{1}_A), \tag{9.2}$$

where $\mathbf{1}_A$ denotes the random variable which is 1 when A occurs and 0 otherwise. This equation holds since $E(\mathbf{1}_A) = 1 \cdot P(\mathbf{1}_A = 1) + 0 \cdot P(\mathbf{1}_A = 0) = P(\mathbf{1}_A = 1) = P(A)$.

Proof of Theorem 9.1. We prove (9.1) by starting with $\operatorname{Var} X = E((X - \mu)^2)$ and changing it in a way does not increase its value.

$$E((X - \mu)^2) \geq E((X - \mu)^2 \, \mathbf{1}_{|X-\mu| \geq \lambda}). \tag{9.3}$$

This inequality holds because $(X - \mu)^2 \geq (X - \mu)^2 \, \mathbf{1}_{|X-\mu| \geq \lambda}$: both sides or equal for all $\omega \in \Omega$ such that $|X(\omega) - \mu| \geq \lambda$, and for all other values of ω, the right hand side is zero.

Similarly, we have

$$(X - \mu)^2 \, \mathbf{1}_{|X-\mu| \geq \lambda} \geq \lambda^2 \mathbf{1}_{|X-\mu| \geq \lambda},$$

because the two sides are $(X(\omega) - \mu)^2$ and λ^2 if $|X(\omega) - \mu| \geq \lambda$, and both sides are zero otherwise. Taking the expectation of both sides, we have

$$E((X - \mu)^2 \, \mathbf{1}_{|X-\mu| \geq \lambda}) \geq E(\lambda^2 \mathbf{1}_{|X-\mu| \geq \lambda}) = \lambda^2 E(\mathbf{1}_{|X-\mu| \geq \lambda}) = \lambda^2 P(|X - \mu| \geq \lambda).$$

Putting this inequality together with (9.3), we get (9.1). □

9.2 The law of large numbers

Sometimes we estimate the expected value of a random variable by using a computer to sample from its distribution 10,000 times and computing the average of these 10,000 samples. This procedure makes sense because the average of 10,000 samples is very likely to be very close to the true mean of the distribution. The law of large numbers formalizes this idea.

Theorem 9.3. (Weak law of large numbers) Let $X_1, X_2 \ldots$, be i.i.d. random variables such that $E(X_1) = \mu$ and $\text{Var}(X) = \sigma^2$ are finite. For $n \geq 1$, define $A_n = (X_1 + \cdots + X_n)/n$. Then for all $\epsilon > 0$, we have

$$P(|A_n - \mu| \geq \epsilon) \to 0 \tag{9.4}$$

as $n \to \infty$.

Remark 9.4. Note that Proposition 6.12 is very similar to the law of large numbers: it says that the mean of A_n is μ, and the variance of A_n decreases to zero as $n \to \infty$. Loosely speaking, this means that the distribution of A_n clusters around μ as $n \to \infty$. Chebyshev's inequality is the tool that helps us bridge the gap between the variance-based notion of clustering in Proposition 6.12 and the probability-based notion in (9.4).

Proof. Combining Proposition 6.12 and Theorem 9.1, we get

$$P(|A_n - \mu| \geq \epsilon) \leq \epsilon^{-2} \text{Var}(A_n) = \sigma^2 \epsilon^{-2} / \sqrt{n} \to 0,$$

as $n \to \infty$. □

Example 9.5. How many times do we need[26] to flip a coin which has probability 60% of turning up heads in order to ensure that the proportion of heads is between 59% and 61% with probability at least 99%?

Solution. We define X_k to be the indicator of the event that the kth flip is heads. Note that $\text{Var}(X_k) = 0.6 \cdot 0.4 = 0.24$. Solving $\sigma^2 \epsilon^{-2} / \sqrt{n} \leq 0.01$ for n and substituting $\sigma^2 = 0.24$ and $\epsilon = 0.01$ tells us that 5.76×10^{10} flips suffices. □

Note that using Chebyshev's inequality here does not give us a complete answer to our question. We know that 5.76×10^{10} flips are enough, but in fact many fewer flips would work. The following Julia code tells us that 15,861 flips will do!

[26] We'll interpret this question to mean "find a number of flips that works" rather than "find the minimum number of flips that would work".

```
binom(n,m,p) = exp(lgamma(n+1) - lgamma(m+1) - lgamma(n-m+1)
                                    + m*log(p) + (n-m) * log(1-p))
function find_least_n()
    p = 0.6
    for n = 1:20_000
        if sum([binom(n,k,p) for k=1+iceil(0.59n):1+ifloor(0.61n)]) > 0.99
            return n
        end
    end
end
find_least_n()
```

A few words of explanation: the function `binom` defined in the first line uses the `lgamma` function, which efficiently calculates the natural logarithm of the factorial of its argument minus 1. We use this function because directly calculating binomial probabilities causes computational overflows, due to the enormous size of binomial coefficients and the tiny size of p^m and $(1-p)^{n-m}$. The function `find_least_n` loops through all the positive integers up to 20,000 and returns the first integer it finds which assigns a probability mass of at least 0.99 to the outcomes between the ceiling[27] of $0.59n$ and the floor of $0.61n$.

Example 9.5 shows that the estimate in Chebyshev's inequality is sometimes exceedingly pessimistic. In the next section, we will develop a much sharper tool.

10 Central limit theorem

Let $X_1, X_2, \ldots$ be sequence of independent, identically distributed random variables, and let $S_n = X_1 + \cdots + X_n$. If we let the X's be Bernoulli distributed with $p = 1/2$ and plot the probability mass functions of $S_1, S_2, S_3, \ldots$, then we notice that the graphs of these pmfs looks increasingly Gaussian-shaped as n gets larger (see Figure 4).

Replacing the Bernoulli distribution with the uniform distribution on $\{1,2,3,4,5,6\}$ gives similar results (Figure 5). The Poisson distribution with mean 5 (Figure 6) also has Gaussian-shaped partial-sum pmfs. Based on these results, we may suspect that the same Gaussian curve emerges as the approximate shape of the distribution of the sum of a large number of independent, identically distributed random variable, regardless of the distribution of the i.i.d. summands.

If we want to obtain a convergence statement corresponding to this observation, we have to account for the fact that the pmfs in Figures 4 through 6 are sliding to the right and spreading out as $n \to \infty$. We do this by forming the *standardized sum* S_n^*, which we define

[27] The names of the functions `iceil` and `ifloor` mean "integer ceiling" and "integer floor".

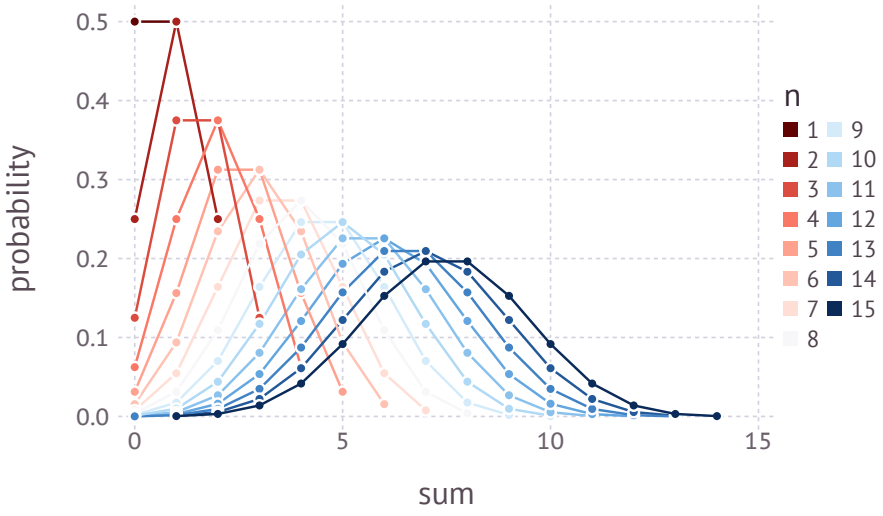

Figure 4: The probability mass functions for the sum $S_n = X_1 + \cdots + X_n$ of n Bernoulli random variables.

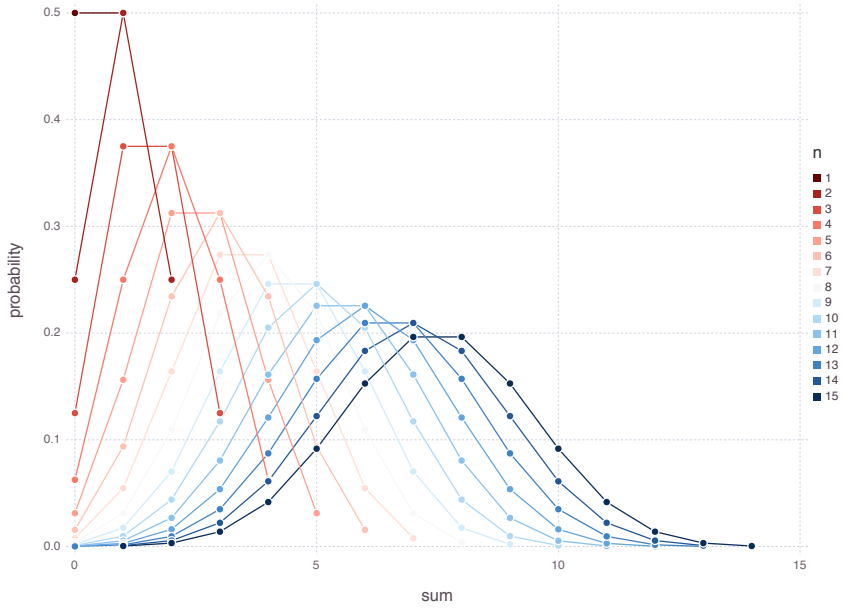

Figure 5: The probability mass functions for the sum $S_n = X_1 + \cdots + X_n$ of n uniform random variables.

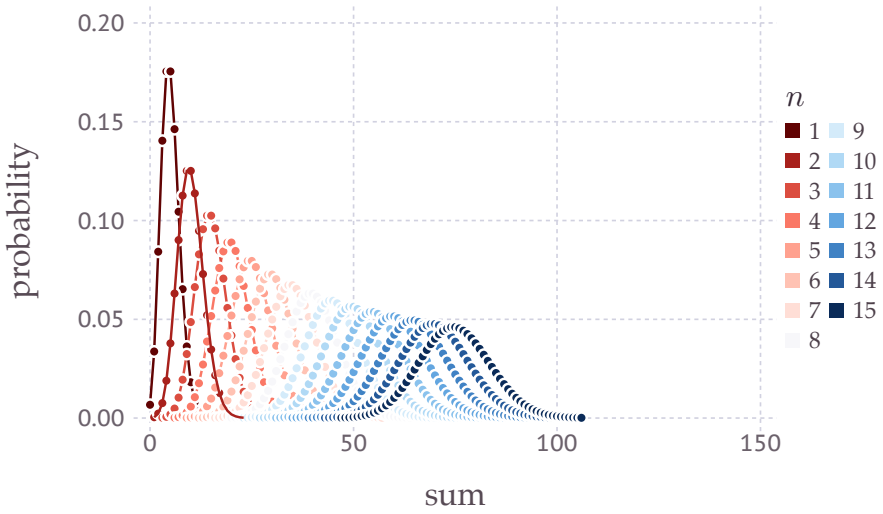

Figure 6: The probability mass functions for the sum $S_n = X_1 + \cdots + X_n$ of n Poisson random variables with mean 5.

as

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}},$$

where μ and σ are the mean and standard deviation of the distribution of X_1.

Exercise 10.1. Show that S_n^* has mean zero and variance 1.

Since we conjectured that S_n is distributed approximately like a random variable with a Gaussian distribution, Exercise 10.1 would lead us to conjecure that S_n^* is distributed approximately like a Gaussian distribution with mean zero and variance 1. The central limit theorem makes this notion precise. Although an investigation of discrete random variables motivated our discussion of the CLT, the theorem is true for continuous random variables too.

Theorem 10.2. (Central limit theorem (CLT)) Let $X_1, X_2, \ldots,$ be i.i.d. discrete or continuous random variables with mean μ and finite standard deviation σ. Then for all $-\infty \leq a < b \leq \infty$,

$$\lim_{n\to\infty} P(a < S_n^* < b) = P(a < Z < b),$$

where $Z \sim \mathcal{N}(0,1)$; in other words, Z has pdf $\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Remark 10.3. Theorem 10.2 remains true if any or all of the strict inequalities are replaced by nonstrict inequalities.

Exercise 10.4. Consider an i.i.d. sum of n Bernoulli random variables with $p = 1/2$, and let $m_n : \mathbb{R} \to [0,1]$ be the pmf of S_n^*. Show that $\lim_{n\to\infty} m_n(x) = 0$ for all $x \in \mathbb{R}$, and explain why this does not contradict Theorem 10.2.

Let's do Example 9.5 again with the CLT.

Example 10.5. How many times do we need to flip a coin which has probability 60% of turning up heads in order to ensure that the proportion of heads is between 59% and 61% with probability at least 99%? Use the central limit theorem to approximate the answer.

Solution. As before, we calculate the standard deviation $\sigma = \sqrt{(0.4)(0.6)}$ and the mean $\mu = 0.6$ of each flip, and we use these values to rewrite the desired probability in terms of S_n^*. We find

$$\begin{aligned} P\left(0.59 < \frac{1}{n}S_n < 0.61\right) &= P\left(-0.01 < \frac{S_n - \mu n}{n} < 0.01\right) \\ &= P\left(-\frac{0.01\sqrt{n}}{\sqrt{0.4 \cdot 0.6}} < \frac{S_n - \mu n}{\sigma\sqrt{n}} < \frac{0.01\sqrt{n}}{\sqrt{0.4 \cdot 0.6}}\right), \end{aligned}$$

where the last step was obtained by multiplying all three expressions in the compound inequality by $\sqrt{n}/\sigma$. Since S_n^* is distributed approximately like a standard normal random variable, the CLT approximation implies that we want to find the least n so that

$$\int_{-a_n}^{a_n} dt > 0.99, \tag{10.1}$$

where $a_n = 0.01\sqrt{n}/\sqrt{0.4 \cdot 0.6}$. By the symmetry of the Gaussian density, we may rewrite (10.1) as

$$\int_{-\infty}^{a_n} dt > 0.995.$$

Defining the normal cdf $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$, we want to find the least integer n such that a_n exceeds $\Phi^{-1}(0.995)$. We can use the Julia code[28]

```
Phiinv(p) = sqrt(2)*erfinv(2*p-1)
Phiinv(0.995)
```

to find that $\Phi^{-1}(0.995) \approx 2.575829$. Setting this equal to a_n and solving for n gives 15,924. Clearly, this approximation is much closer to the true value of 15,861 than the bound 5.76×10^{10} we got from Chebyshev's inequality. □

In the next section we will see how to prove the central limit theorem.

[28] The error function `erf` and its inverse `erfinv` are more often available in computing systems the normal cdf Φ and its inverse Φ. The error function is defined by $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$ and is related to Φ by $\Phi(x) = \frac{1}{2}(1 + \operatorname{erf}(x/\sqrt{2}))$. This implies that $\Phi^{-1}(p) = \sqrt{2}\operatorname{erf}^{-1}(2p - 1)$.

11 Moment generating functions

11.1 Introduction

Let X be a random variable. For all integers $n \geq 0$, we define the nth moment $\mu_n = E(X^n)$, assuming $E(|X|^n) < \infty$. The zeroth moment of any random variable is 1, while the first moment is the mean. The second moment is related to the variance:

$$\text{Var}(X) = E(X^2) - \mu^2 \implies \mu_2 = \mu^2 + \text{Var}(X).$$

The key idea in the present section is to package *all* of the moments of a random variable into a single mathematical object. We form a power series with coefficients $\mu_k/k!$ as k goes from 0 to ∞:

$$g(t) = \mu_0 + \mu_1 t + \frac{\mu_2}{2}t^2 + \frac{\mu_3}{6}t^3 + \cdots.$$

So far this might not seem particularly helpful: it seems like we'd have to do infinitely many calculations to find all the moments in order to calculate g. However, if we did know g somehow, then we could recover all the moments by taking derivatives of g and substituting $t = 0$:

Exercise 11.1. Show that $\mu_k = g^{(k)}(0)$, where $g^{(k)}$ denotes the kth derivative of g.

Here's where the magic happens: we write μ_n as $E(X_n)$ and then reverse-distributing the E (using linearity[29]), we have

$$g(t) = E\left(1 + tX + \frac{1}{2}(tX)^2 + \frac{1}{6}(tX)^3 + \cdots\right),$$

and, recognizing the Taylor series for e^{tX}, we get

$$g(t) = E(e^{tX}).$$

So, although g was defined using the moments of X, we can actually find g directly by calculating $E(e^{tX})$ and then come up with all the moments of X by differentiating the result.

Example 11.2. Use mgfs to find the mean and variance of a binomial random variable with parameters n and p.

Solution. Let $X \sim \text{Bin}(n, p)$ and calculate

$$g(t) = E(e^{tX}) = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{n}\binom{n}{k}(pe^t)^k(1-p)^{n-k} = (pe^t + (1-p))^n.$$

[29] We're fudging a bit here, since linearity applies to finite sums and this is an infinite one. However, this step can be rigorously justified.

Differentiating gives

$$g'(t) = \left(pe^t - p + 1\right)^{n-1} npe^t,$$

which when $t = 0$ equals np. Differentiating again gives

$$g''(t) = \left(pe^t - p + 1\right)^{n-2}(n-1)np^2e^{2t} + \left(pe^t - p + 1\right)^{n-1} npe^t$$

which equals $\mu_2 = (np - p + 1)np$ when $t = 0$. Subtracting $\mu_1^2 = (np)^2$ from μ_2 gives a variance of $np(1-p)$. □

Exercise 11.3. Suppose that the mgf of X is g. Show that the mgf of $X + a$ is $e^{at}g(t)$, and the mgf of bX is $g(bt)$. Show that if Y is independent of X and has mgf $\tilde{g}$, then the mgf of $X + Y$ is $g(t)\tilde{g}(t)$.

11.2 Uniqueness

The moment generating function of a distribution is defined in terms of its moments, which are in turn defined in terms of the distribution itself. Can this procedure be reversed? In other words, if we know the mgf of a distribution, is that enough to say what the distribution is? Roughly speaking, the answer to the question is yes, as long as the moments don't grow too fast. The following theorem is proved in Billingsley's *Probability and Measure*, Chapter 30.

Theorem 11.4. Let X be a random variable all of whose moments $\mu_0, \mu_1, \mu_2, \ldots$ are finite, and suppose that there exists $\epsilon > 0$ so that the power series

$$g(t) = \sum_{k=0}^{\infty} \frac{\mu_k}{k!} t^k$$

converges for all t between $-\epsilon$ and ϵ. Then the law of X is uniquely determined by $g(t)$. In other words, if Y is another random variable whose mgf is equal to g, then Y and X have the same law.

11.3 Probability generating function

If a random variable X is supported on the nonnegative real numbers, then its law is characterized by the sequence $p_0, p_1, \ldots$, where $p_k = P(X = k)$ for all $k \geq 0$. Thus

$$g(t) = \sum_{k=0}^{\infty} e^{tk} p_k,$$

which we can rewrite using the substitution $z = e^t$ as

$$h(z) = \sum_{k=0}^{\infty} p_k z^k.$$

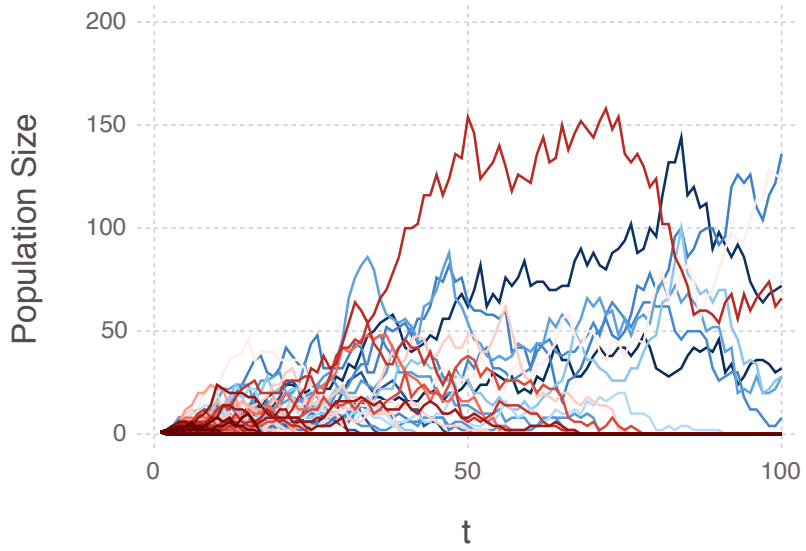

Figure 7: Population versus time for 500 trials of a Galton-Watson process with $p_0 = p_2 = 1/2$.

This function h is called the *probability generating function* of the sequence $p_0, p_1, p_2, \ldots$, since probabilities rather than moments are used as the power series coefficients (additionally, there is no factorial factor in this case).

Exercise 11.5. Show that $h'(1) = E(X)$.

11.4 Branching processes

Fix a probability distribution P supported on the nonnegative integers, specified as the preceding section by the probabilities $p_0, p_1, p_2 \ldots$ assigned to $0, 1, 2, \ldots$. Consider a population which starts with one individual who at the end of one time step perishes and leaves X offspring behind, where X is a random variable with law P. In the next time step, each member of the population again perishes and leaves a random number of offspring behind according to the law P. All these offspring random variables are independent, within each time interval and across time intervals. This is called a *Galton-Watson* process.

For example, if $p_0 = 1/2$ and $p_1 = 1/2$, then the population survives for a Geom$(1/2)$ amount of time and then goes extinct. If $p_1 = 1$, then the population size is constant for all time, and if $p_2 = 2$, then the population grows exponentially over time. These examples show that, depending on the measure P, it is possible to achieve certain extinction or certain non-extinction.

Other measures are make for a more interesting Galton-Watson process. For example, suppose $p_0 = p_2 = 1/2$. Population size as a function of time are shown in Figure 7 for 500 trials of this process. The following Julia code simulates this process:

```
population = [1] # we'll grow this list with the population values
for j=1:n-1
    # Calculate the number of offspring
    # as 2*rand(Bool) for each current population
    # member, and append the result to the list
    # of population values:
    push!(population,sum(2*rand(Bool,vals[end])))
end
```

Most of these processes terminate by time 100, but 8 of them survived. Is the probability of nonextinction positive? That is the question we'll answer in this section.

We begin by defining T to be the extinction time (with $T = \infty$ if there is no extinction), and $d_m = P(T \leq m)$ to be the probability that extinction occurs by the mth time step. We can decompose the probability space according to the number of offpsring on the first step. Denote by X_1 the population size after one time step. Then

$$\begin{aligned} d_m = P(T \leq m) &= P\left(\bigcup_{j \geq 0} \{T \leq m\} \cap \{X_1 = j\}\right) \qquad (11.1) \\ &= P(X_1 = 0) + P(X_1 = 1)d_{m-1} + P(X_1 = 2)d_{m-1}^2 + P(X_1 = 3)d_{m-1}^3 + \cdots, \end{aligned}$$

where we have used the fact that if there are j offspring on the first step, then the probability of extinction by time m is d_{m-1}^j. This is true because each of the j populations which descend from each of the j offspring at time 1 must all go extinct in the next $m-1$ time steps. So (11.1) gives

$$d_m = p_0 + p_1 d_{m-1} + p_2 d_{m-1}^2 + p_3 d_{m-1}^3 + \cdots \qquad (11.2)$$

Since d_m is an increasing sequence which is bounded above by 1, there exists some real number $d \leq 1$ such that d_m converges to d. Taking $m \to \infty$ on both sides of (11.2), we get

$$d = p_0 + p_1 d + p_2 d^2 + \cdots = h(d),$$

where h is the probability generating function of $p_0, p_1, \ldots$. Since $\sum_{k=1}^{\infty} p_k = 1$, we see that $d = 1$ is a solution of $d = h(d)$. Furthermore,

$$h'(z) = \sum_{k=1}^{\infty} k p_k z^{k-1},$$

and

$$h''(z) = \sum_{k=2}^{\infty} k(k-1) p_k z^{k-2},$$

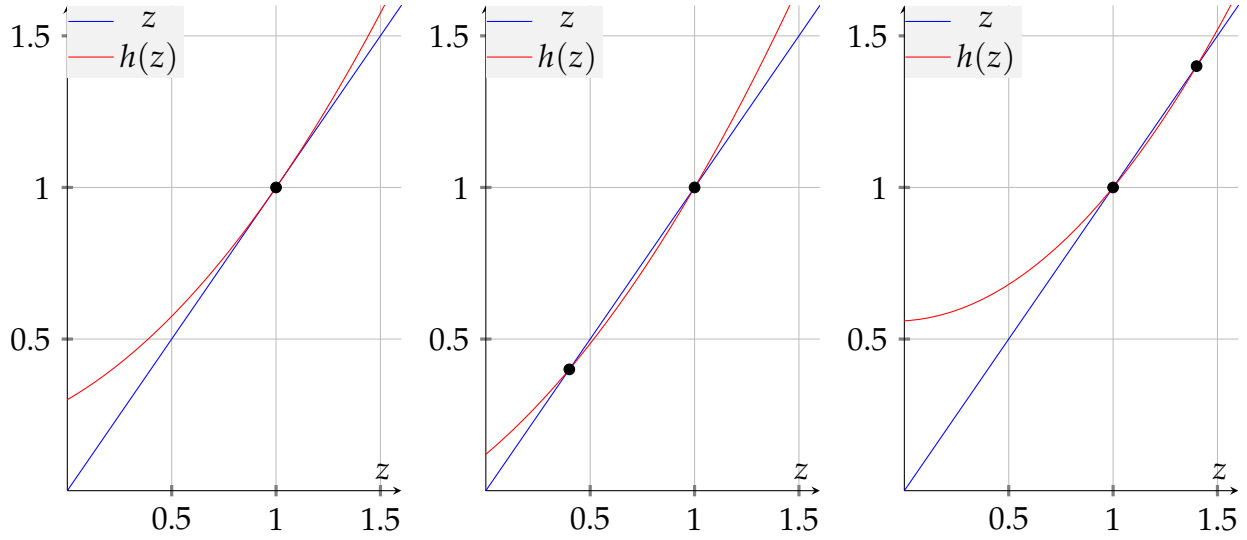

Figure 8: The three possibilities for the intersection points of $h(z)$ and z, given that h is an increasing convex function whose graph includes the point $(1,1)$: (1) the graphs are tangent, and the only intersection point is $z = 1$, (2) the graphs are not tangent, and the second intersection point is less than 1, and (3) the graphs are not tangent, and the second intersection point is greater than 1.

which shows that $h'(z)$ and $h''(z)$ are both nonnegative whenever z is nonnegative. So if we consider the graph of the identity function z and the function $h(z)$, they intersect at $z = 1$, and h is an increasing, convex function[30]. Thus there are three possibilities for the graphs of $h(z)$ and z: (1) the graph of h is tangent to the graph of the identity function at $z = 1$, (2) the graph of h intersects the graph of the identity function at $z = 1$ and at one other point between 0 and 1, and (3) the graphs intersect at $z = 1$ and at one other point greater than 1 (see Figure 8).

One way to distinguish these cases is to consider $h'(1)$. If $h'(1) = 1$, then the graphs are tangent. If $h'(1) > 1$, then the other intersection point is less than 1, and if $h'(1) < 1$, then the other intersection point is greater than 1. By Exercise 11.5, $h'(1)$ is equal to the mean of the offspring distribution, so already we have a partial answer to our question: if the mean number of offspring is less than or equal to 1, then there are no solutions to the equation $h(d) = d$ which are strictly less than 1, so the process goes extinct with probability 1.

If the mean of the offspring distribution is greater than 1, then there exists a solution less than 1. The following theorem says that this solution is indeed equal to d.

Theorem 11.6. Suppose that $p_k > 0$ for some value of $k \geq 2$. If $h'(1) \leq 1$, then the Galton-Watson process goes extinct with probability 1. Otherwise, the process goes extinct with

[30] If the only positive p values are p_0 and p_1, then the process goes extinct with probability 1 unless $p_1 = 1$, in which case the process never goes extinct. So let's assume that $p_k > 0$ for some $k \geq 2$, so that $h''(z) > 0$ for all $z > 0$.

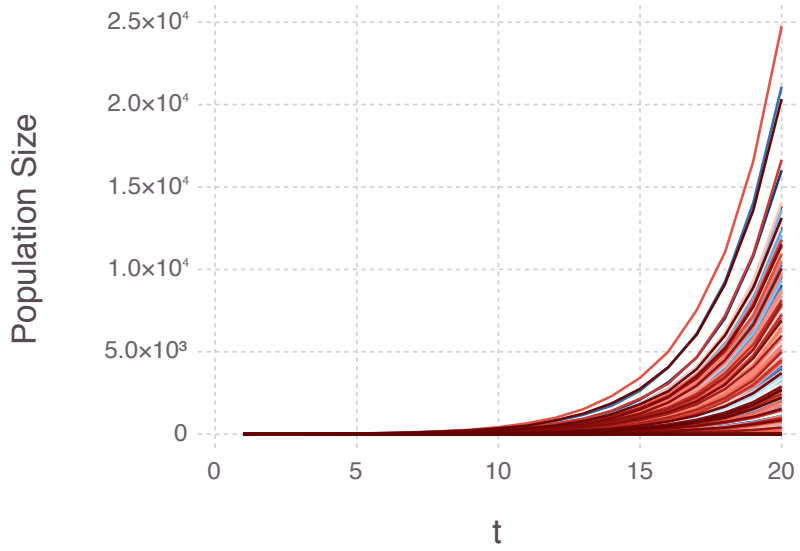

Figure 9: Population versus time for 1000 trials of a Galton-Watson process, with $p_0 = p_3 = 1/2$. Compare with the critical Galton-Watson process in Figure 7.

probability equal to the smallest solution of the equation $h(z) = z$, which is strictly less than 1.

Proof. We have already seen that the process goes extinct if $h'(1) \leq 1$ and that the least root r of $h(z) = z$ is less than 1 if $h'(1) > 1$. It remains to show that $d = r$.

Recall that $d_1 = p_0 = h(0)$ and $d_m = h(d_{m-1})$ for $m \geq 1$. Note that for $z < r$, we have $h(z) < r$, since otherwise h would be nonincreasing over the interval (z, r), in contradiction with the fact that $h'(z) > 0$ for all $z > 0$. Thus the sequence $d_0, h(d_0), h(h(d_0)), \ldots$ is bounded above by r. Since $d \in \{1, r\}$ and d is not greater than r, we conclude $d = r$. □

We call a Galton-Watson process with $h'(1) = 1$ *critical*, a process with $h'(1) < 1$ is *subcritical*, and a process with $h'(1) > 1$ is *supercritical*. See Figure 9 to see how differently supercritical and critical processes behave.

11.5 Proof of the central limit theorem

In this section, we aim to explain why the Gaussian distribution emerges as the universal limiting distribution in the central limit theorem.

Exercise 11.7. Show that the moment generating function of $\mathcal{N}(0, 1)$ is $e^{t^2/2}$. (Hint: set

up the integral to compute $E(e^{tX})$, and multiply and divide by $e^{t^2/2}$ so as to complete the square on the inside of the integral.)

We will need one other theorem, whose proof is rather advanced and will be omitted. See Billingsley's *Convergence of probability measures* to see a proof. It says that pointwise convergence of mgfs implies pointwise convergence of the corresponding cdfs.

Theorem 11.8. Suppose that $(F_n)_{n=1}^{\infty}$ is a sequence of cumulative distribution functions corresponding to probablity meausres on $\mathbb{R}$ whose moment generation functions are defined on an open interval containing the origin and are denoted $(g_n)_{n=1}^{\infty}$. If g is an mgf such that $g_n(t) \to g(t)$ for all t in an open interval containing 0, and if F is the cdf corresponding to g, then $F_n(x) \to F(x)$ at every point where F is continuous.

Now we can prove the central limit theorem.

Theorem 11.9. Suppose that $(X_n)_{n=1}^{\infty}$ is an i.i.d. sequence of random variables, and suppose that the moment generating function of X_1 is defined in an open interval. Denote by F_n the cdf of $S_n^* = (X_1 + \cdots + X_n - n\mu)/(\sigma\sqrt{n})$. Then for all $x \in \mathbb{R}$, we have

$$F_n(x) \to \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}}\, du,$$

as $n \to \infty$.

Proof. First, it suffices to consider the case $\mu = 0$, since we can apply the mean-zero version to $X_1 - \mu, X_2 - \mu, \ldots$ to obtain the more general version. So assume $\mu = 0$.

In light of Theorem 11.8 and Theorem 11.4, it suffices to calculate the mgf g_n of S_n^* and show that

$$\lim_{n\to\infty} g_n(t) = e^{t^2/2},$$

since the mgf of $\mathcal{N}(0,1)$ is $e^{t^2/2}$, by Exercise 11.7. The mgf of $X_1/(\sigma\sqrt{n})$ is given by $g\left(\frac{t}{\sigma\sqrt{n}}\right)$, by Exercise 11.3. Since the X_k's are independent, then, we get

$$g_n(t) = \left[g\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$

Taking the log of both sides and defining $G_n(t) = \log g_n(t)$ and $G(t) = \log g(t)$, we get

$$G_n(t) = n\, G\left(\frac{t}{\sigma\sqrt{n}}\right).$$

Taylor's theorem with remainder tells us that

$$G(u) = G(0) + G'(0)u + \frac{1}{2}G''(0)u^2 + \frac{1}{6}G'''(\xi)u^3,$$

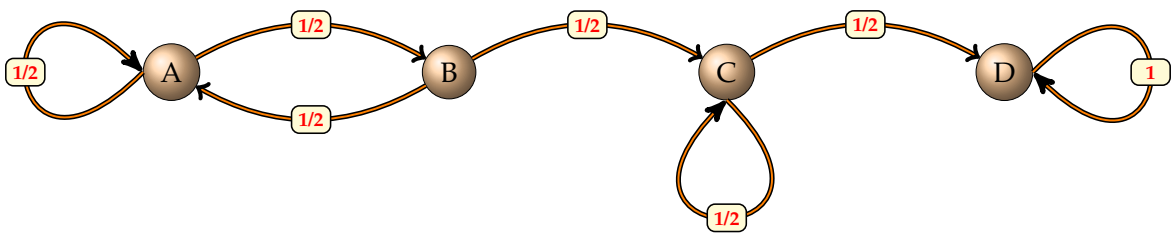

Figure 10: An example of a state space and transition probabilities for a Markov chain: a random walker moves around on this graph with probabilities specified by the arrow labels.

where ξ is some number between 0 and u. By direct calculation, we get that $G(0) = G'(0) = 0$, while $G''(0) = \sigma^2$. Thus

$$G_n(t) = n\left[0 + 0 + \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + \frac{1}{6}G'''(\xi)\left(\frac{t}{\sigma\sqrt{n}}\right)^3\right]$$
$$= \frac{t^2}{2} + \frac{1}{6}G'''(\xi)\left(\frac{t}{\sigma}\right)^3\frac{1}{\sqrt{n}}.$$

Since ξ is between 0 and $t/(\sigma\sqrt{n})$ and since G''' is continuous[31], the third term goes to 0 as $n \to \infty$. Exponentiating both sides (and using the continuity of the exponential function), we get $g_n(t) \to \exp(t^2/2)$ for all $t \in \mathbb{R}$, as desired. □

12 Markov chains

12.1 Introduction

Consider a walker which moves around on the graph in Figure 10, at each step moving from its current location to a new one along one of the outgoing arrows. The probabilities for these moves are specified by the arrows' labels, and the moves are determined by independent coin flips (unless the walker is at D, in which case the walker stays at D forever). The vertices A, B, C, and D are called *states*, and the arrow labels are called *transition probabilities*. The sequence of states visited by the walker is an example of a *Markov chain*.

Figure 11 shows 100 trials of this process, all started from A.

[31] ...since a convergent power series is infinitely differentiable; this is a basic result in analysis.

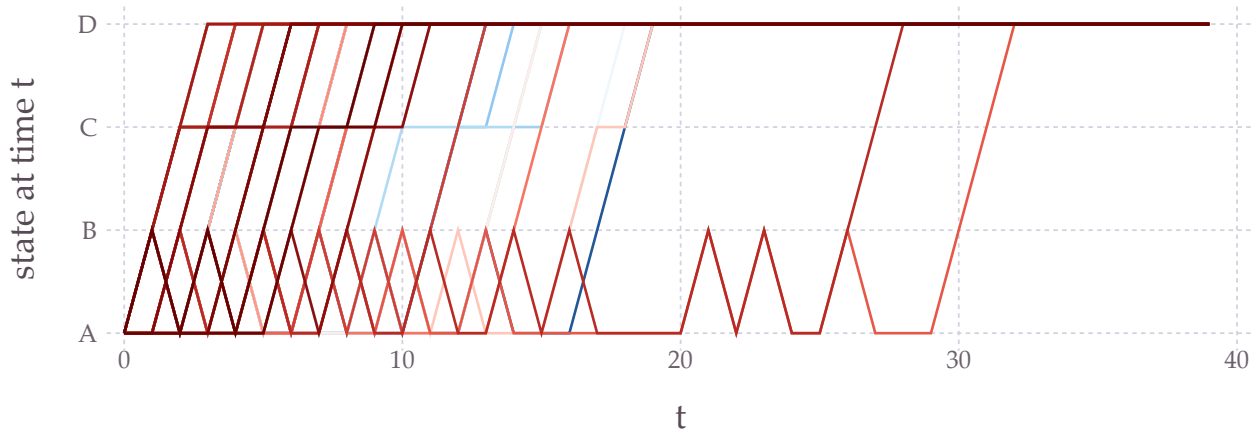

Figure 11: One hundred trials of the Markov chain shown in Figure 10.

Question 12.1. Here are three questions raised by Figure 11.

(i) What is the probability that the walker ends up in state A after, say, 10 steps?
(ii) Does the walker eventually end up in state D, with probability 1?
(iii) What is the expected value of the number of times that state A is visited?

FIgure 11 suggests an affirmative answer to Question 12.1(ii), while (i) and (iii) are less straightforward. In this section, we will develop systematic methods for answering all these questions.

12.2 Transition matrix

We begin by giving a precise definition of a Markov chain. Let S be a finite set, and let $(X_t)_{t=0}^{\infty}$ be a sequence of S-valued random variables (also called an S-valued *process*). We refer to S as the *state space* and its elements as *states*.

Definition 12.2. The process $(X_t)_{t=0}^{\infty}$ is said to be a *Markov chain* if there exists a matrix[32] $\mathbf{P} = (P_{i,j})_{i,j \in S}$ so that for all $i_0, i_1, \ldots, i_{t-2}, i, j \in S$, we have

$$P(X_t = j \,|\, X_0 = i_0 \text{ and } X_1 = i_1 \text{ and } \cdots \text{ and } X_{t-1} = i) = P_{i,j}.$$

Loosely speaking, a sequence of S-valued random variables is a Markov chain if the conditional probabilities for what happens on the next step are the same regardless of whether we condition on the whole history of the process or just on the current state.

Note that this definition makes no reference to the starting state X_0. It will often be some

[32]This notation means that $\mathbf{P}$ is a matrix whose rows and columns are indexed by elements of S. If i and j are states, then $P_{i,j}$ denotes the (i,j)th entry of the matrix.

fixed element of S, but it can also be random. We will generally be interested in properties of Markov chains which do not depend on the starting state. For this reason, a Markov chain is generally regarded as being fully specified by its transition matrix $\mathbf{P}$.

We have already seen an example of a Markov chain in Figure 10. Here's an example of a process which is *not* a Markov chain:

Example 12.3. Consider an i.i.d. sequence of dice rolls $(Y_t)_{t\geq 0}$, and define X_t to be the mode[33] of $Y_0, \ldots, Y_t$ (break ties with precedence). For example, if the dice rolls were

$$2, 3, 3, 1, 4, 5, 4, 6, 4, \ldots,$$

then the values of the the process $(X_t)_{t=0}^{\infty}$ would be $2, 2, 3, 3, 3, 3, 3, 3, 4, \ldots$.

This process is not a Markov chain because[34], if (for example) 3 just recently became the mode, then whichever number was previously the mode has a decent chance of overtaking 3 on the next step. If, instead, the process has been stuck at 3 for a long time, it's very likely that the next X_t is also equal to 3. So, we see that the whole history of the process provides meaningful information about the probabilities for what happens next.

The matrix $\mathbf{P}$ in Definition 12.2 is called the *transition matrix* of the Markov chain. For example, the transition matrix of the Markov chain in Figure 10 is

$$\mathbf{P} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}.$$

Clearly a transition matrix must contain nonnegative entries, and its rows must sum to 1. For any square matrix $\mathbf{P}$ satisfying these two conditions, there is[35] a Markov chain whose transition matrix is $\mathbf{P}$.

The next proposition shows why it's handy to represent the transition probabilities in matrix form:

Proposition 12.4. For $t \geq 0$ and $i, j \in S$, define $P_{i,j}^{(t)}$ to be the probability that $X_t = j$ given that $X_0 = i$. Define $\mathbf{P}^t$ to be the tth power of the matrix $\mathbf{P}$, and denote by $(\mathbf{P}^t)_{i,j}$ the (i, j)th entry of $\mathbf{P}^t$. Then for all $t \geq 0$ and $i, j \in S$, we have

$$P_{i,j}^{(t)} = (\mathbf{P}^t)_{i,j}.$$

[33] The number which appears most often.

[34] If this intuitive explanation is unsatisfactory, you can prove that this process is not Markov by noting that X_1 is always equal to X_0, whereas X_2 has a positive probability of being different from X_1.

[35] There isn't much to prove here: one defines the Markov chain inductively using $\mathbf{P}$ to define the transition probabilities.

Proof. We explain the $t = 2$ case; the general case follows similarly, by induction. We decompose the probability space according to the value of X_1:

$$P(X_2 = j \mid X_0 = i) = \sum_{\ell \in S} P(X_2 = j \text{ and } X_1 = \ell \mid X_0 = i).$$

By the Markov property, this sum equals

$$\sum_{\ell \in S} P(X_2 = j \mid X_1 = \ell) P(X_1 = \ell \mid X_0 = i) = \sum_{\ell \in S} P_{\ell,j} P_{i,\ell} = \sum_{\ell \in S} P_{i,\ell} P_{\ell,j}.$$

In words, this formula says that $P_{i,j}^{(2)}$ is equal to the dot product of the ith row of $\mathbf{P}$ and the jth column of $\mathbf{P}$. By definition of the matrix product, this is equal to $(\mathbf{P}^2)_{i,j}$. □

Exercise 12.5. Suppose that the initial state X_0 is chosen according to the probability measure represented by the $1 \times n$ vector v. (For example, if X_0 is uniformly distributed in S, then v would be the column vector with entries $(1/n, 1/n, \ldots, 1/n)$, where n is the number of elements of S. If $X_0 = 2$ with probability 1, then v would be $(0, 1, 0, 0, \ldots, 0)$, etc.) Then the distribution of X_t is given by $v\mathbf{P}^t$.

Proposition 12.4 answers Question 12.1 (i). The Julia code

```
P = [1//2 1//2   0    0  ;
     1//2  0    1//2  0  ;
      0    0    1//2 1//2;
      0    0     0    1  ]
P^10
```

returns the 10th power of $\mathbf{P}$:

$$\mathbf{P}^{10} = \begin{bmatrix} \frac{89}{1024} & \frac{55}{1024} & \frac{88}{1024} & \frac{792}{1024} \\ \frac{55}{1024} & \frac{34}{1024} & \frac{55}{1024} & \frac{880}{1024} \\ 0 & 0 & \frac{1}{1024} & \frac{1023}{1024} \\ 0 & 0 & 0 & \frac{1024}{1024} \end{bmatrix}.$$

So the probability of being in state A after 10 steps is 89/1024. Proposition 12.4 also suggests a way of addressing Question 12.1 (ii): note that most of the probability mass in the first row has shifted to the last entry. If we separate off the last row and column, then we can write $\mathbf{P}$ as

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & 1 \end{array}\right],$$

where $\mathbf{Q}$ is the 3×3 matrix containing the transition probabilities between states A, B, and C, and $\mathbf{R}$ is the 3×1 matrix of transition probabilities from A, B, and C to D. The

final row consists of the 1×3 zero matrix $\mathbf{0}$ followed by a 1. Doing the block matrix multiplication[36], we see that

$$\mathbf{P}^n = \left[\begin{array}{c|c} \mathbf{Q}^n & * \\ \hline \mathbf{0} & 1 \end{array}\right],$$

where the asterisk denotes some expression involving $\mathbf{Q}$ and $\mathbf{R}$ whose value we don't care about. If we can show that $(\mathbf{Q}^n)_{i,j} \to 0$ for all $i, j \in S$ as $n \to \infty$, then that will establish that the probability of ending up in state D tends to 1 as $n \to \infty$. This may be done directly in this case using linear algebra[37], but in the next section we will develop an approach which is both more general and more probabilistic.

12.3 Absorbing Markov chains

State D in the Markov chain in Figure 10 is *absorbing*, meaning that $P_{\mathrm{D,D}} = 1$. A state i with $P_{i,i} < 1$ will be called *non-absorbing*.

Definition 12.6. An *absorbing Markov chain* is a Markov chain with at least one absorbing state which also satisfies the property that from any non-absorbing state, it is possible to transition to an absorbing state in some number of steps (in other words, for every non-absorbing state i, there exists an absorbing state j and an integer t so that $P_{i,j}^{(t)} > 0$).

If $\mathbf{P}$ is a transition matrix for an absorbing Markov chain, then obtain the *canonical form* of $\mathbf{P}$ by arranging the states so that the m non-absorbing states come before the n absorbing states. If we divide the matrix into four submatrices by drawing horizontal and vertical lines separating non-absorbing and absorbing states, then we get

$$\mathbf{P} = \left[\begin{array}{c|c} \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right],$$

where $\mathbf{Q}$ is an $m \times m$ matrix of transition probabilites between absorbing states, $\mathbf{R}$ is an $m \times n$ matrix of transition probabilites from non-absorbing to absorbing states, and $\mathbf{I}$ is an $n \times n$ identity matrix.

Proposition 12.7. Let $\mathbf{P}$ be a transition matrix for an absorbing Markov chain. Then[38] $\lim_{t\to\infty} \mathbf{Q}^t = \mathbf{0}$. Furthermore, this convergence happens at a geometric rate, meaning that there exists $\lambda > 1$ so that $\lim_{t\to\infty} \mathbf{Q}^t \lambda^t = \mathbf{0}$.

[36] This is an idea from linear algebra. It says that if you decompose a matrix into blocks in this way, then matrix multiplication may be carried out with the blocks by treating them as if they were matrix entries.

[37] If you've taken a linear algebra course: the absolute values of the eigenvalues of $\mathbf{Q}$ are all less than 1, which implies via diagonalizing $\mathbf{Q}$ that $\mathbf{Q}^n \to 0$.

[38] Convergence to the zero matrix means that the (i, j)th entry converges to 0, for all i, j.

Proof. For each non-absorbing state i, define r_i to be the minimum number of moves such that the probability p_i of transitioning from i to some absorbing state in r_i moves is positive. Define r to be the largest of the r_i values and p to be the smallest of the p_i values, so that after r moves, the probability of being in an absorbing state is at least p, regardless of starting position.

Then the probability of being in a non-absorbing state after r moves is at most $1 - p$, the probability of being in a non-absorbing state after $2r$ moves is at most $(1 - p)^2$, and so on. Since $(1 - p)^k \to 0$ as $k \to \infty$, the probability of being in an absorbing state after kr moves tends to 0 as $k \to \infty$. Writing $t = kr$ and setting $\lambda = (1 - p/2)^{-1/r}$, we see that $\lambda^t \mathbf{Q}^t \to 0$ as well. □

The motivation for writing $\mathbf{P}$ in block form is that the answer to some fundamental questions about the behavior of the chain can be expressed in terms of $\mathbf{Q}$ and $\mathbf{R}$. The following theorem gives a prime example.

Theorem 12.8. The matrix[39] $\mathbf{I} - \mathbf{Q}$ is invertible, and for non-absorbing states i and j, the (i, j)th entry of its inverse is equal to the expected amount of time spent at j for a Markov chain started at i.

Proof. We write the number of visits to j as a sum of indicator random variables: let Z_t be 1 if $X_t = j$ and let Z_t be 0 otherwise. Then the number of visits to j is equal to $\sum_{t=0}^{\infty} Z_t$. Taking the expectation of this random variable gives

$$\sum_{t=0}^{\infty} E(Z_t) = \sum_{t=0}^{\infty} P(X_t = j) = \sum_{t=0}^{\infty} \mathbf{Q}_{i,j}^t.$$

By Proposition 12.4, this is equal to the (i, j)th entry of the matrix

$$\sum_{t=0}^{\infty} \mathbf{Q}^t.$$

This sum converges by Proposition 12.7 (apply the comparison test entry-by-entry) to a matrix $\mathbf{N}$ with finite entries.

We can calculate $\mathbf{N}$ in the same way we calculate the sum of a geometric series of numbers. The identity

$$(\mathbf{I} - \mathbf{Q})(\mathbf{I} + \mathbf{Q} + \mathbf{Q}^2 + \cdots + \mathbf{Q}^{n-1}) = \mathbf{I} - \mathbf{Q}^n$$

may be seen by distributing and collecting terms on the left-hand side. Taking $n \to \infty$, we get

$$(\mathbf{I} - \mathbf{Q})\mathbf{N} = \mathbf{I}.$$

Thus $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$. □

[39] This $\mathbf{I}$ is different from the one that appears in block form of $\mathbf{P}$, since its dimensions must match those of $\mathbf{Q}$ in order for matrix subtraction to make sense. In general, $\mathbf{I}$ may be taken to mean the identity matrix of whatever dimension is implied by the context.

Now we can answer Question 12.1. The Julia code

```
P = [1//2 1//2   0    0  ;
     1//2  0    1//2  0  ;
      0    0    1//2 1//2;
      0    0     0    1  ]
Q = P[1:3,1:3] # select the top-left 3 x 3 block
inv(eye(3) - Q) # eye(3) is the 3 x 3 identity matrix
```

returns

$$(\mathbf{I}-\mathbf{Q})^{-1} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \end{array} \begin{array}{c} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \end{array} \\ \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix} \end{array}$$

As a sanity check, note that we could have calculated the last row without Theorem 12.8. The number of visits to A and B starting from C is clearly 0. The number of visits to C is equal to a geometric random variable with $p = 1/2$, which has expected value 2.

The following exercise provides another example of an important property of a Markov chain can be expressed in terms of $\mathbf{Q}$ and $\mathbf{R}$.

Exercise 12.9. Let i be a nonabsorbing state and j an absorbing state. Show that the probability that a Markov chain started at i ends up in state j is equal to the (i, j)th entry of the matrix $\mathbf{NR}$. (Hint: write the desired probability as $\sum_{t\geq 0}\sum_{\ell\in S} Q_{i,\ell}^{(t)} R_{\ell,j}$.)

12.4 Regular Markov chains

Suppose that we modify the Markov chain from Figure 10 by adjusting one arrow: rather than making D an absorbing state, we assign a transition probability of 1 for the arrow from D to A (see Figure 12). Since there are no absorbing states in this new transition matrix, the resulting Markov chain is not absorbing. In fact, this Markov chain is *regular*, which means that some power of the transition matrix has no zero entries:

Definition 12.10. A transition matrix $\mathbf{P}$ is *regular* if there exists an integer $t > 0$ such that for every pair of states i and j, we have $P_{i,j}^{(t)} > 0$.

Exercise 12.11. Find a value of t satisfying Definition 12.10 for the transition matrix depicted in Figure 12. Hint: you can reason about this directly using Figure 12, doing any matrix calculations.

To understand how this change affects the long term behavior of the Markov chain, let's calculate the 100th power of P: running

```
P = [1//2 1//2   0    0  ;
```

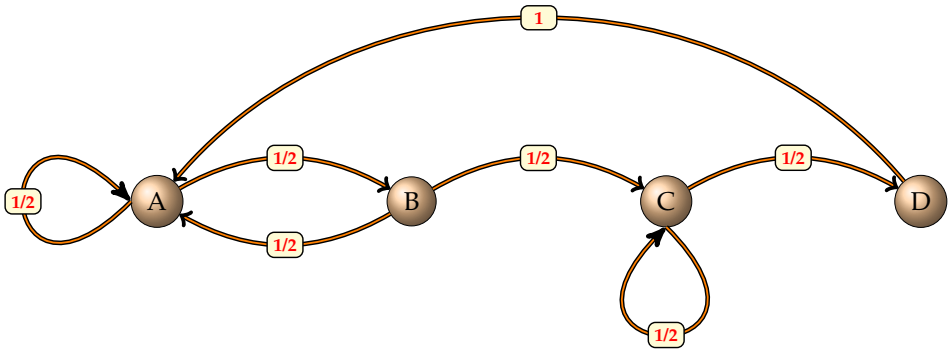

Figure 12: This Markov chain is *regular*, meaning that it is possible to transition from any state to any other state.

```
     1//2  0    1//2  0  ;
      0    0    1//2 1//2;
      1    0     0    0  ]
# we have to convert P to floating point
# to avoid an overflow with rational
# arithmetic
float(P)^100
```

we get

$$\mathbf{P}^{100} \approx \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{B} & \text{C} & \text{D} \end{array} \\ \begin{bmatrix} 0.444444 & 0.222222 & 0.222222 & 0.111111 \\ 0.444444 & 0.222222 & 0.222222 & 0.111111 \\ 0.444444 & 0.222222 & 0.222222 & 0.111111 \\ 0.444444 & 0.222222 & 0.222222 & 0.111111 \end{bmatrix} \end{array}.$$

Rather than concentrating in a single state, the distribution of the location of the walker after many steps is spread out across the four states, apparently converging to the measure which assigns mass $4/9, 2/9, 2/9$, and $1/9$ to states A, B, C, and D, regardless of the starting state. The vector $(4/9, 2/9, 2/9, 1/9)$ is called the *stationary distribution* of this regular Markov chain, and in Theorem 12.15 below we will show that indeed for any regular transition matrix $\mathbf{P}$, the matrix powers $\mathbf{P}^t$ converge to a matrix whose rows are all equal to the stationary distribution.

Let's begin by finding a more direct way to describe the stationary distribution. If we assume that there exists a vector v such that $\mathbf{P}^t \to \mathbf{P}^\infty$, where $\mathbf{P}^\infty$ is a matrix all of whose rows are v, then we have

$$\mathbf{P}^\infty = \lim_{t\to\infty} \mathbf{P}^t = \lim_{t\to\infty} \mathbf{P}^{t+1} = \lim_{t\to\infty} \mathbf{P}^t\mathbf{P} = \mathbf{P}^\infty\mathbf{P},$$

Looking at this matrix identity row by row, it says that $\nu = \nu\mathbf{P}$. In linear algebra terms, we say that ν is a *left eigenvector* of $\mathbf{P}$ with eigenvalue 1.

We can find the stationary vector ν for the example in Figure 12 by solving the equation $\nu = \nu\mathbf{P}$. Subtracting ν from both sides, the equation becomes

$$\nu(\mathbf{I} - \mathbf{P}) = 0. \tag{12.1}$$

Writing ν as $(\nu_1, \nu_2, \nu_3, \nu_4)$ and writing out (12.1) gives

$$\begin{pmatrix} \frac{1}{2}\nu_1 - \frac{1}{2}\nu_2 - \nu_4 & -\frac{1}{2}\nu_1 + \nu_2 & -\frac{1}{2}\nu_2 + \frac{1}{2}\nu_3 & -\frac{1}{2}\nu_3 + \nu_4 \end{pmatrix} = (0 \quad 0 \quad 0 \quad 0).$$

Solving this system along with $\nu_1 + \nu_2 + \nu_3 + \nu_4 = 1$ gives $\nu = (4/9, 2/9, 2/9, 1/9)$, which agrees with that we guessed by raising P to the 100th power.

The following theorem, whose proof may be found starting on p. 448 in the book, ensures us that there is a unique vector satisfying this equation. It is a special case of the *Perron-Frobenius* theorem in linear algebra.

Theorem 12.12. Let $\mathbf{P}$ be a regular transition matrix. Then there exists a unique vector ν with positive entries summing to 1 such that $\nu = \nu\mathbf{P}$.

Given the existence of the stationary vector ν, we can sketch a proof of Theorem 12.15, the *fundamental theorem for regular Markov chains*. We begin with a couple of exercises.

Exercise 12.13. Show that if $(X_t)_{t=0}^{\infty}$ is a Markov chain with state space S, transition matrix $\mathbf{P}$, and stationary distribution ν, and if the initial state X_0 is randomly selected according to the distribution ν, then for all $t \geq 0$, $P(X_t = j) = \nu_j$ for all $j \in S$.

Exercise 12.14. Let (Ω, P) be a probability space, and let A and B be events. Show that

$$P(A) - P(B^c) \leq P(A \cap B) \leq P(A)$$

Conclude that if $(A_n)_{n=1}^{\infty}$ and $(B_n)_{n=1}^{\infty}$ are sequences of events such that $P(A_n) \to p \in [0,1]$ and $P(B_n) \to 1$ as $n \to \infty$ then $P(A_n \cap B_n) \to p$ as $n \to \infty$.

The following proof is due to Doeblin.

Theorem 12.15. As $t \to \infty$, $\mathbf{P}^t$ converges to the matrix $\mathbf{P}^{\infty}$, all of whose rows are equal to the stationary distribution of $\mathbf{P}$.

Proof idea. Probabilistically, this theorem says that if $(X_t)_{t=0}^{\infty}$ is a Markov chain with transition matrix $\mathbf{P}$, then $P(X_t = j) \to \nu_j$ as $t \to \infty$, regardless of the initial state X_0.

To prove this, we define two Markov chains as follows: let $(X_t)_{t=0}^{\infty}$ be a Markov chain with transition matrix $\mathbf{P}$ starting at a state randomly selected according to the stationary distribution ν. By Exercise 12.13, the distribution of X_t is equal to ν for all t.

Consider an arbitrary state $i \in S$. Loosely speaking, we will define a Markov chain $Y = (Y_t)_{t=0}^{\infty}$ which starts from i and evolves independenly of X until the first time Y occupies the same state at the same time as X, after which it moves in lock step with X.

More precisely, we define $Y_0 = i$. For $t \geq 1$, we define Y_t to be equal to X_t if $Y_{t-1} = X_{t-1}$. If $Y_{t-1} \neq X_{t-1}$ then we define Y_t in a way which is independent of the process X and so that the transition probabilities of the process Y are given by $\mathbf{P}$.

Since $\mathbf{P}$ is regular, the probability that X and Y have merged by time t converges to 1 as $t \to \infty$[40]. On the event that X and Y have merged by time t, the distributions of X_t and Y_t are exactly the same, since $X_t = Y_t$ in that case. More precisely, we have

$$P(Y_t = j) = P(Y_t = j \text{ and } X_t = Y_t) + P(Y_t = j \text{ and } X_t \neq Y_t).$$

The second term goes to 0 since $P(X_t \neq Y_t) \to 0$, and the first term goes to ν_j as $t \to \infty$, by Exercise 12.14, since it is equal to $P(X_t = j \text{ and } X_t = Y_t)$. Thus $P(Y_t = j) \to \nu_j$ as $t \to \infty$. □

See the book for more examples of regular Markov chains.

13 Simple symmetric random walks on $\mathbb{Z}^d$

13.1 Introduction

Fix an integer $d \geq 1$, and let $\mathbb{Z}^d$ be the integer grid in d dimensions (that is, the set of all d-tuples of integers). If x and y are elements of $\mathbb{Z}^d$ such that the distance between x and y is 1, we say that x and y are *neighbors*.

Let $X = (X_t)_{t=1}^{\infty}$ be a sequence of random elements of $\mathbb{Z}^d$ determined by setting[41] $X_0 = \mathbf{0}$ and choosing X_t uniformly at random from among the neighbors of X_{t-1}, for all $t \geq 1$. We call X a *simple symmetric random walk*. See Figure 13 for some examples. The focus of the present discussion will be the following question.

Question 13.1. Does X visit the origin infinitely many times with probability 1, or is there a positive probability that X visits the origin only finitely many times?

We begin with the following observation.

[40] To see this, let j be some state and r some integer so that the $P_{i,j}^{(r)} > 0$ for all $i \in S$. Let $p > 0$ be a lower bound for all the values $P_{i,j}^{(r)}$ as i ranges over S. Then the probability that X and Y haven't merged after kr steps is at most $(1-p)^k$, which tends to 0 as $k \to \infty$.

[41] We will use the boldface $\mathbf{0}$ to denote the zero vector in $\mathbb{Z}^d$.

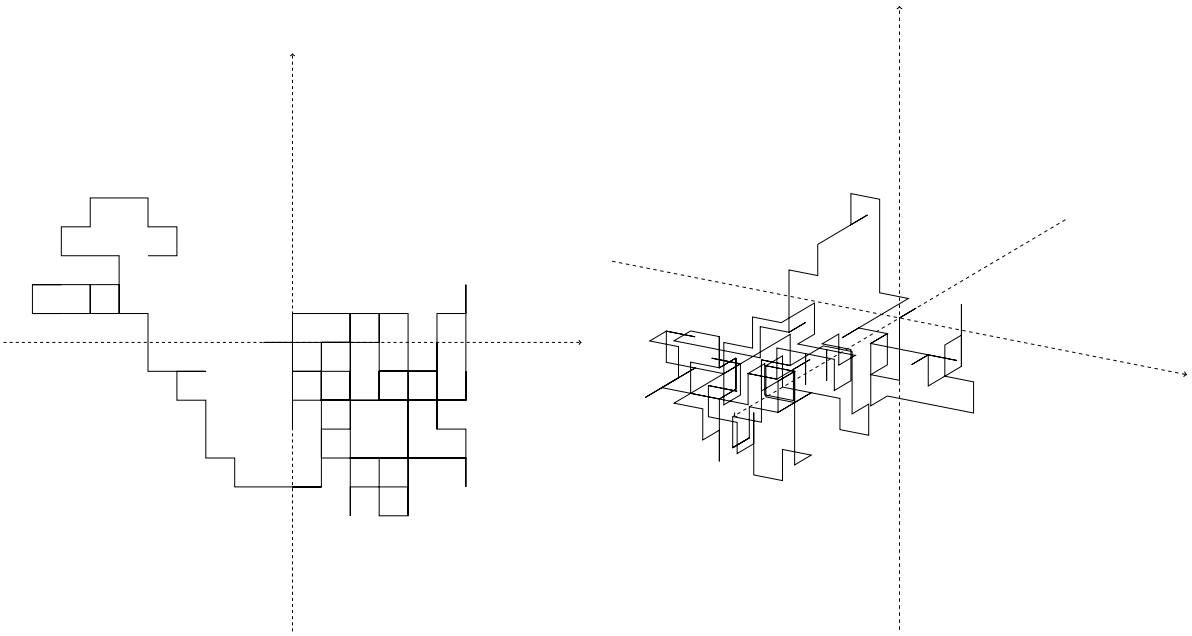

Figure 13: Sample runs of the first 200 steps of simple symmetric random walks in two and three dimensions.

Proposition 13.2. The number of visits to the origin is either infinite almost surely, or it is geometric with a success probability strictly betwen 0 and 1.

Proof. Let p be the probability that X visits the origin at least once. Then $p > 0$, since the probability is $\frac{1}{2d}$ that X revisits the origin on the second step. So $p \in (0, 1]$.

We can view the portions of the random walk between the visits to the origin as a Bernoulli trials process, with success probability p. More precisely, starting from the origin, we let the random walk run until it visits the origin again (which we count as a success) or drifts off to infinity without ever visiting the origin again (which we count as a failure). So the number of visits to the origin is geometrically distributed with success probability p. If $p = 1$, this means that that the number of visits to the origin is infinite almost surely. □

Definition 13.3. If X visits the origin finitely many times with probability 1, then X is said to be *transient*. If X visits the origin infinitely many times with probability 1, X is said to be *recurrent*.

The following corollary of Proposition 13.2 follows from the fact that every geometrically distributed random variable has finite expected value. Denote by V the number of times that X visits the origin.

Corollary 13.4. Every simple symmetric random walk is either transient or recurrent. Furthermore, the random walk is transient if $E(V) < \infty$ and recurrent if $E(V) = \infty$.

In light of Corollary 13.4, we turn our attention to calculating the expected number of visits to the origin. We write the expected number of visits as a sum of indicator random variables. We let I_t equal 1 if $X_t = \mathbf{0}$ and $I_t = 0$ otherwise. Then $V = \sum_{t=0}^{\infty} I_t$, so

$$E(V) = \sum_{t=1}^{\infty} P(X_t = \mathbf{0}).$$

Theorem 13.5. The simple symmetric random walk on $\mathbb{Z}^1$ is recurrent.

Proof. We note that $P(X_t = \mathbf{0}) = 0$ if t is odd and

$$P(X_t = \mathbf{0}) = \binom{t}{t/2} \left(\frac{1}{2}\right)^t$$

if t is even. Applying Stirling's formula to the right-hand side, we get

$$P(X_t = \mathbf{0}) \asymp \frac{1}{\sqrt{t}}$$

for even values of t. Since $\sum_{t \text{ odd}} \frac{1}{\sqrt{t}} = \infty$, the result follows from Corollary 13.4. □

Theorem 13.6. The simple symmetric random walk on $\mathbb{Z}^2$ is recurrent.

Proof. Directly calculating $P(X_t = (0,0))$ is more difficult in two dimensions, but we can use a trick[42] to make it easier. We define A_t and B_t to be $\sqrt{2}$ times the projections of X onto the lines $y = x$ and $y = -x$, respectively. Then A_t and B_t evolve simultaneously as simple symmetric random walks, and $X_t = (0,0)$ if and only if both A_t and B_t are zero. Furthermore, the steps of A_t and B_t are independent (exercise).

So $P(X_t = (0,0)) = P(A_t = B_t = 0) \asymp (1/\sqrt{t})^2 = 1/t$ when t is even, and $P(X_t = (0,0)) = 0$ otherwise. Since $\sum_{t \text{ even}} \frac{1}{t}$ diverges, X is recurrent. □

We conclude this section by showing that the simple symmetric random walk on $\mathbb{Z}^3$ is transient.

Theorem 13.7. The simple symmetric random walk on $\mathbb{Z}^3$ is transient.

Proof. Note that $X_t = (0,0,0)$ if and only if there exist nonnegative integers i, j, and k summing to n such that X has taken i positive and i negative steps in the z-direction, as well as j positive and j negative steps in the y-direction, and k positive and k negative

[42] Much of the remainder of this section draws from the notes at http://www.statslab.cam.ac.uk/~james/Markov/s16.pdf.

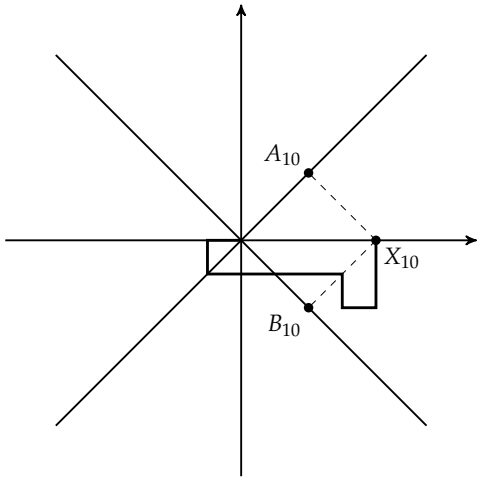

Figure 14: We define the random walks A_t and B_t to be the projections of X_t onto the diagonal lines.

steps in the x-direction. This implies that $P(X_t = (0,0,0)) = 0$ if t is odd. If $t = 2n$ for some integer n, then $P(X_t = (0,0,0))$ is equal to

$$\sum_{i+j+k=n} \frac{(2n)!}{(i!j!k!)^2} \left(\frac{1}{6}\right)^{2n} = \frac{(2n)!}{n!} \cdot \frac{1}{2^{2n}3^n} \sum_{i+j+k=n} \frac{n!}{(i!j!k!)^2} \left(\frac{1}{3}\right)^n.$$

Now if n is a multiple of 3, then $i!\,j!\,k!$ is minimized when $i = j = k = n/3$ (exercise). So the expression above is bounded above by

$$\frac{(2n)!}{n!} \cdot \frac{1}{2^{2n}3^n} \cdot \frac{1}{((n/3)!)^3} \sum_{i+j+k=n} \frac{n!}{i!j!k!} \left(\frac{1}{3}\right)^n.$$

The sum is now equal to 1, because it is the total mass of the probability measure associated with placing n balls uniformly and independently into 3 bins. Applying Stirling's formula then tells us that the above formula is asymptotic to $n^{-3/2}$.

If n is k less than a multiple of 3, where $k \in \{2,4\}$, then $P(X_{2n} = (0,0,0)) \leq 6^{k/2} P(X_{2n+k} = (0,0,0))$, since $X_{2n} = (0,0,0)$ implies that $X_{2n+2} = (0,0,0)$ on the event of probability $1/6$ on which $X_{2n+2} - X_{2n+1} = -(X_{2n+1} - X_{2n})$, and similarly for $k = 4$. Thus, we lose only a constant factor by ignoring the values of t which are not multiples of 6.

Since $\sum_{n=1}^{\infty} n^{-3/2} < \infty$, we conclude that X is transient. □

Exercise 13.8. Show that the simple symmetric random walk on $\mathbb{Z}^d$ is transient for all $d \geq 3$. (Hint: use the $d = 3$ case.)

Exercise 13.9. Consider the random walk whose steps are i.i.d. $\{-1,1\}$-valued random variables with probability $p \in (0,1)$ associated to each right step and $1-p$ to each left step. Show that X is transient for all $p \neq 1/2$.