Math 531 Probability Theory
Lecture Notes

Version January 14, 2023

Timo Seppäläinen

Benedek Valkó

NOT FOR DISTRIBUTION!

Contents

Warm-up

Probability theory deals with the 'likelihood' of uncertain events, the outcomes of random experiments. In Chapter 1 we introduce the precise framework needed to work in probability theory, and we get to know some of the basic tools.

In the current chapter (which may be skipped) we discuss a couple of examples where we can compute probabilities without really knowing the precise definition, based only on some symmetry properties. We also discuss an example that shows why a precise framework is necessary.

Uniformly chosen elements from a finite set

Suppose that we have a random experiment with finitely many outcomes where because of some symmetry (either by construction or by assumption) we know that the outcomes are equally likely. If the set of outcomes is denoted by Ω then by symmetry the probability of each particular outcome should be $\frac{1}{\#\Omega}$. (Here $\#\Omega$ denotes the size of the set Ω.) Moreover, the probability that the outcome is in a specific subset A of Ω should be $\frac{\#A}{\#\Omega}$.

Example 0.1. If we flip an unbiased coin then we can assume that the two possible outcomes (heads and tails) are equally likely, hence they both have probability $\frac{1}{2}$.

If we flip an unbiased coin 10 times then an outcome is a length 10 sequence of heads or tails. There are 2^{10} such sequences, we may assume that they are equally likely. Thus the probability of getting 10 tails in a row is 2^{-10}. △

Example 0.2. Suppose that we flip an unbiased coin until we get heads. What is the probability that we need at least five coin flips for that?
We have to flip the coin at least five times exactly when we get tails for the first four coin flips. The probability of this is 2^{-4} by the previous example. (Note, that the random experiment in question does not have equally likely outcomes, but we could compute the probability in question by making a comparison to another experiment.) △

Example 0.3 (Birthday problem)**.** Suppose that we invite n 'randomly chosen' guests to a party. What is the probability that there will be at least two guests with the same birthday?

We have to make some modeling choices in order to solve this problem. We disregard leap years, and assume that all possible combination of birthdays are equally likely. We may assume $2 \leq n \leq 365$.

Let us encode each possible birthday with an integer between 1 and 365. When we record the birthdays of each guest then we get a sequence of length n built from these integers. There are 365^n such sequences, and each one is equally likely by our assumption, so the probability of seeing a specific sequence of birthdays is 365^{-n}. To find the probability of getting at least two matching birthdays we just need to count how many of these sequences will have at least two identical elements. It is easier to count those sequences where all the elements are different, we have $365 \cdot 364 \cdots (366 - n)$ of these. Thus there are exactly $365^n - 365 \cdot 364 \cdots (366 - n)$ sequences with at least two numbers being equal, and hence the probability in question is

$$\frac{365^n - 365 \cdot 364 \cdots (366 - n)}{365^n} = 1 - \prod_{k=0}^{n-1} (1 - \tfrac{k}{365}).$$

One can check that for $n = 22$ this is approximately 0.4757 and for $n = 23$ it is approximately 0.5073. For $n = 50$ it is approximately 0.9704.

If repeated approximation of the formula is tedious, we can get a good estimate using the following trick. For small x we have $1 - x \approx e^{-x}$ (this follows from the Taylor expansion of e^x near 0). Hence, as long as $\frac{n-1}{365}$ is not too large, we have

$$1 - \prod_{k=0}^{n-1} (1 - \tfrac{k}{365}) \approx 1 - \prod_{k=1}^{n-1} e^{-\frac{k}{365}} = 1 - e^{-\sum_{k=1}^{n-1} \frac{k}{365}} = 1 - e^{-\frac{n(n-1)}{2 \cdot 365}} \approx 1 - e^{-\frac{n^2}{2 \cdot 365}}.$$

The resulting approximation is pretty close to the real values for $n = 22, 23, 50$. By keeping track of the error terms one could get precise upper and lower bounds using the previous idea. △

Uniform element from a region of finite size

Suppose now that we have a random experiment where the outcome is a randomly chosen point from a subset Ω of $\mathbb{R}^d$ (for some fixed d). Moreover, assume that Ω has finite size (length/area/volume), and that the point is chosen 'uniformly' (its not clear what that means, but let us go with the flow...). Then it is reasonable to conclude that the probability that the outcome is in a specific subset A of Ω should be $\frac{|A|}{|\Omega|}$. (Here $|\cdot|$ is the appropriate d-dimensional volume.) This is just an extension of the previously discussed discrete setup.

Example 0.4. Suppose that we choose a point randomly (and uniformly) from a unit square. What is the probability that the chosen point is closer than 1/4 to at least one of the sides?

The points in the unit square that are at least of distance 1/4 from all sides are in a square of side 1/2 (with sides parallel to the unit square, and having the same center). Thus the area of the set of points which are closer than 1/4 to at least one of the sides is equal to $1 - (\frac{1}{2})^2 = \frac{3}{4}$, and this also gives the probability in question. △

The next example describes a famous random experiment due to Georges-Louis Leclerc, Comte de Buffon, from the 18th century.

Example 0.5. We have a grid of horizontal parallel lines which are unit distance apart. We throw down a needle with length $\ell < 1$ from a significant height, so that the position of the needle is 'random'. What is the probability that it will intersect one of the lines if we assume that 'all positions' are equally likely?

In order to check that the needle intersects one of the lines it is enough to know the following two parameters: the distance d of the midpoint of the needle to the nearest grid line below it and the angle α of the needle measured from the horizontal direction.

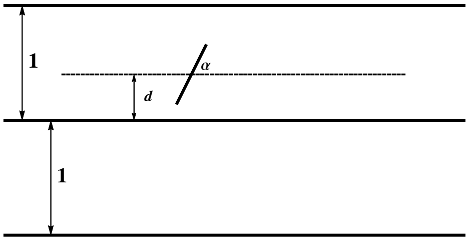

Figure 1. Buffon's needle.

Since 'all positions' are equally likely, it is reasonable to assume that the point (d, α) is a 'uniformly' chosen point from the rectangle $[0,1] \times [0,\pi]$. The needle will intersect one of the grid lines if and only if $\frac{1}{2}\ell \sin\alpha \geq \min(d, 1-d)$. After some integration we can check that the area of the region

$$A = \{(d,\alpha) : d \in [0,1), \alpha \in [0,\pi), \tfrac{1}{2}\ell \sin\alpha \geq \min(d, 1-d)\}$$

is given by 2ℓ. Hence the probability in question is $\frac{2\ell}{\pi}$. △

We finish with another famous example considered by Joseph Bertrand in the late 19th century.

Example 0.6 (Bertrand's paradox)**.** Consider a unit circle and inscribe an equilateral triangle. Now choose a chord of the circle 'randomly'. What is the probability that this random chord is at least as long as the side of the triangle?

Bertrand gave three different solutions giving three different answers.

1. Suppose we choose two random points on the circle and connect them. If we assume that all configurations are equally likely then we might as well fix one of the points (since we only care about the length of the chord). Then the result of the experiment is an angle in $[0, 2\pi)$, and if we assume that the fixed point corresponds to the angle 0 then the outcomes which produce a large enough chord are the angles in $[2\pi/3, 4\pi/3]$. Hence the probability is $\frac{2\pi/3}{2\pi} = \frac{1}{3}$.

2. The length of the chord only depends on the distance of its midpoint from the center of the circle. Suppose that we just choose this distance uniformly from $[0, 1]$. The chord will be at least as large as the side of the triangle if the distance is at least $1/2$, which has a probability $1/2$.

3. Suppose that we know choose a point uniformly from the unit disk, and draw the chord which has this point as its midpoint. A quick computation with areas gives a probability $1/4$.

So which one is the correct answer? The three solutions describe three different ways to choose a randomly chosen chord. Since the original problem did not define the 'randomness' carefully enough, each one of these solutions could be correct. This shows that in general 'randomly choosing something' can mean lots of different things, and we have to be careful with the definitions. △

Chapter 1

Events and their probabilities

This chapter introduces the basic concepts of the mathematical theory of probability, in terms of which all the subsequent discussion is conducted. These are the sample spaces, events, probability measures, conditional probability, and independence.

1.1. Probability spaces

Probability theory begins by setting down the axioms of a mathematical structure that can model an experiment with random outcomes and quantify the probabilities of different results of the experiment. Before stating the formal definition, we illustrate through an example the conventions and ingredients of a mathematical model of random outcomes.

Example 1.1. Consider the experiment of rolling a fair six-sided die. The outcome of the experiment is one of the numbers in the set $\{1, 2, 3, 4, 5, 6\}$. We express this in mathematical terms by defining the *sample space* of this experiment as $\Omega = \{1, 2, 3, 4, 5, 6\}$. By convention, a total probability of one is divided among the outcomes. The assumption that the die is *fair* means that all outcomes are equally likely, and hence receive an equal proportion of the total probability. Thus the *probability* of each outcome $\omega \in \Omega$ is given by $P\{\omega\} = \frac{1}{6}$.

Additionally, we want to know the probabilities of more complicated scenarios. For example, what is the probability that the outcome is larger than two? This is an example of an *event*, and is mathematically represented by a subset of Ω. If we denote this event by A, we may express it in any of the following equivalent ways:

$$A = \{\text{the outcome is larger than two}\} = \{\omega \in \Omega : \omega > 2\} = \{3, 4, 5, 6\}.$$

It is a fundamental property of the theory that the probability of an event is the sum of the probabilities of its constituent pieces:

$$P(A) = P\{3, 4, 5, 6\} = P\{3\} + P\{4\} + P\{5\} + P\{6\} = 4 \cdot \tfrac{1}{6} = \tfrac{2}{3}.$$

△

These are Kolmogorov's axioms for probability theory.

Definition 1.2. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ with the following three components.

(a) Ω is a set, called the **sample space**.

(b) $\mathcal{F}$ is a collection of subsets of Ω. Members of $\mathcal{F}$ are called **events**. $\mathcal{F}$ is assumed to be a σ**-algebra**, which means that it satisfies the following three properties.
   (b.1) $\Omega \in \mathcal{F}$. That is, the whole sample space itself is an event.
   (b.2) If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.
   (b.3) If $\{A_k\}_{1 \le k < \infty}$ is a sequence of members of $\mathcal{F}$, then their union $\bigcup_{k=1}^{\infty} A_k$ is also a member of $\mathcal{F}$.

(c) P is a function from $\mathcal{F}$ into real numbers, called the **probability measure**. P satisfies the following axioms.
   (c.1) $0 \le P(A) \le 1$ for each event $A \in \mathcal{F}$.
   (c.2) $P(\varnothing) = 0$ and $P(\Omega) = 1$.
   (c.3) If $\{A_k\}_{1 \le k < \infty}$ is a sequence of pairwise disjoint events then

$$P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k). \tag{1.1}$$

△

The purpose of a probability space $(\Omega, \mathcal{F}, P)$ is to model an experiment with random outcomes. The sample space Ω (upper case Greek letter omega) is the set of all the possible outcomes of the experiment. Elements of Ω are called *sample points* and typically denoted by ω (lower case omega). The sample points in Example 1.1 are the integers 1 through 6.

Events are the subsets of Ω for which a probability can be given. Technically speaking, the word *collection* is a synonym of the word *set*. It is used in (b) above simply to avoid the repetitive phrase "set of subsets". A synonym for σ-algebra that also appears in the literature is σ*-field*.

For each event $A \in \mathcal{F}$, the number $P(A) \in [0,1]$ given by the probability measure P is the *probability of the event* A. $\varnothing$ is the empty set, that is, the subset of Ω that contains no sample points. *Pairwise disjoint* means that $A_k \cap A_\ell = \varnothing$ for each pair of distinct indices $k \ne \ell$. An equivalent phrase is that the events $\{A_k\}$ are *mutually exclusive*. Axiom (c.3) can be paraphrased in English by saying that the probability of a union of countably many mutually exclusive events is equal to the sum of their probabilities.

Note that the additivity rule (c.3) applies also to finitely many events. Namely, if $A_1, A_2, \dots, A_n$ are pairwise disjoint events then

$$P(A_1 \cup \cdots \cup A_n) = P(A_1) + \cdots + P(A_n). \tag{1.2}$$

This is a special case of (1.1) where $A_k = \varnothing$ for all $k > n$. Property (1.2) is *finite additivity* while property (1.1) is the axiom of *countable additivity*.

Important remark. The precise definition of $\mathcal{F}$ as a σ-algebra was included in Definition 1.2 mainly for the sake of completeness. The notion of a σ-algebra does not play a role in this book, though we may occasionally refer to it. This is not because σ-algebras are not important. On the contrary, σ-algebras are central to probability theory because they represent information. But working with σ-algebras requires measure theory which is beyond the prerequisites for this book. The good news is that a great deal of interesting, useful and deep probability theory can be covered without σ-algebras. △

Often we are interested in some numerical value associated to the outcome of a random experiment. This just means that we are interested in the value of a function that maps the elements of the sample space into the real numbers. These functions are called random variables.

Definition 1.3. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A **random variable** on Ω is a real valued function $X : \Omega \to \mathbb{R}$, for which $\{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F}$ for any $c \in \mathbb{R}$. ♣

The additional condition on the function ensures that we can actually work with the function X in our probability space, in the sense that we can answer questions like *"what is the probability that $X \leq 3$?"* or *"what is the probability that X is equal to 5?"*. In these notes we will most often see functions on Ω that automatically satisfy the second condition. In fact, it is very hard to come up with a natural example for a function $\Omega \to \mathbb{R}$ in a probability space, that is not a random variable. We will have a closer look at random variables in the upcoming chapters.

Our first task is to get used to applying the axioms to construct probability spaces and to calculate probabilities for events of interest. This is achieved by going through examples. Just as sets in general, sample spaces come in three types: finite, countably infinite, and uncountable. We illustrate each with examples.

Example 1.4 (Continuation of Example 1.1)**.** In Example 1.1 we can calculate the probability of each subset $A \subset \Omega$ by counting points:

$$P(A) = \sum_{\omega:\omega\in A} P\{\omega\} = (\#A) \cdot \tfrac{1}{6}. \tag{1.3}$$

Above $\#A$ denotes the cardinality of A, that is, the number of points in A. Since each subset has a well-defined probability, we can take as the class of events $\mathcal{F}$ the *power set* of Ω, that is, the collection of all subsets of Ω. Here are three notational ways of expressing this:

$$\mathcal{F} = \{A : A \subset \Omega\} = 2^{\Omega} = \mathscr{P}(\Omega).$$

If the die we roll is not fair, the probabilities have to be altered to reflect our expectations. If a three is twice as likely as any other number, we use the probability measure $\widetilde{P}$ defined by

$$\widetilde{P}\{1\} = \widetilde{P}\{2\} = \widetilde{P}\{4\} = \widetilde{P}\{5\} = \widetilde{P}\{6\} = \tfrac{1}{7} \quad \text{and} \quad \widetilde{P}\{3\} = \tfrac{2}{7}.$$

For a third variant, suppose we scratch away the four from the original fair die and turn it into a second three. Then the probability measure to use is

$$Q\{1\} = \tfrac{1}{6},\ Q\{2\} = \tfrac{1}{6},\ Q\{3\} = \tfrac{2}{6},\ Q\{4\} = 0,\ Q\{5\} = \tfrac{1}{6},\ Q\{6\} = \tfrac{1}{6}.$$

Since three different probability measures are discussed in the same example, we distinguished them notationally from each other by using P, $\widetilde{P}$ and Q. This type of notational variation in the service of clarity is typical in mathematics. △

Equation (1.3) illustrates a general rule: whenever all outcomes are equally likely,

$$P(A) = \frac{\#A}{\#\Omega}. \tag{1.4}$$

This formula is expressed by saying that the probability of A is the number of favorable outcomes divided by the total number of outcomes.

Probability measures $\widetilde{P}$ and Q in the previous example illustrate that outcomes do not have to be equally likely, and some portions of the sample space can have probability zero.

When an event consists of a single sample point, as for example $\{1\}$ in the example above, notational consistency would demand writing $P(\{1\})$ for the probability of the event $\{1\}$. We can simplify this to $P\{1\}$ or $P(1)$ without any harm. However, we should still keep in mind that according to the axioms the probability measure P is a function on $\mathcal{F}$, the collection of events, and not on the sample space Ω.

Cartesian product spaces arise naturally as sample spaces because probability frequently deals with repetitions of a simple experiment, such as the roll of a die. If $A_1, A_2, \dots, A_n$ are sets then their Cartesian product

$$\prod_{i=1}^{n} A_i = A_1 \times A_2 \times \cdots \times A_n$$

is by definition the set of ordered n-tuples with the ith element from A_i:

$$A_1 \times A_2 \times \cdots \times A_n = \{(x_1, \dots, x_n) : x_i \in A_i \text{ for } i = 1, \dots, n\}.$$

The cardinality of a Cartesian product is calculated by multiplying:

$$\#(A_1 \times A_2 \times \cdots \times A_n) = (\#A_1) \cdot (\#A_2) \cdots (\#A_n).$$

Example 1.5. Roll a fair die three times and record the numbers in the order in which they appear. Each outcome of the experiment is an ordered triple $\omega = (s_1, s_2, s_3)$ where s_i is the result of the ith roll. Thus

$$\begin{aligned}\Omega &= \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} \\ &= \{\omega = (s_1, s_2, s_3) : s_i \in \{1,2,3,4,5,6\} \text{ for each } i = 1,2,3\}.\end{aligned}$$

With a fair die each outcome is equally likely. So, since $\#\Omega = 6^3 = 216$, the probability of each individual outcome is $P\{\omega\} = 6^{-3} = \frac{1}{216}$. As before, $\mathcal{F}$ is the power set of Ω. With equally likely outcomes the probability of an event A is computed from

$$P(A) = \sum_{\omega : \omega \in A} P\{\omega\} = \frac{\#A}{\#\Omega}.$$

As an example, let us calculate probability that the first roll is a one or two and the other two rolls are even:

$$\begin{aligned}&P\{\text{the first roll is a one or two and the other two rolls are even}\} \\ &= P\{\omega = (s_1, s_2, s_3) : s_1 \in \{1,2\}, s_2, s_3 \in \{2,4,6\}\} \\ &= \frac{\#(\{1,2\} \times \{2,4,6\} \times \{2,4,6\})}{\#\Omega} = \frac{2 \cdot 3 \cdot 3}{216} = \frac{1}{12}.\end{aligned}$$

This example generalizes naturally. For N rolls of a fair die each outcome is an ordered N-tuple $\omega = (s_1, s_2, \dots, s_N)$ of integers 1 through 6. The sample space is

$$\begin{aligned}\Omega &= \{1,2,3,4,5,6\}^N \\ &= \{\omega = (s_1, s_2, \dots, s_N) : s_i \in \{1,2,3,4,5,6\} \text{ for each } i = 1,2,\dots,N\}.\end{aligned} \tag{1.5}$$

$P\{\omega\} = 6^{-N}$ for each sample point $\omega \in \Omega$.

Here are a couple of examples for random variables on this probability space:
- the sum of the three numbers on the three dies: $X(s_1, s_2, s_3) = s_1 + s_2 + s_3$,
- the value of the second die: $Y(s_1, s_2, s_3) = s_2$,
- the maximum of the three numbers: $Z(s_1, s_2, s_3) = \max(s_1, s_2, s_3)$.

Because $\mathcal{F}$ contains all subsets of Ω, and function $\Omega \to \mathbb{R}$ is a random variable. △

Example 1.6. We flip a fair coin 5 times. Encoding heads as 0 and tails as 1 the outcome of the experiment is a sequence $(s_1, s_2, s_3, s_4, s_5)$ with $a_i \in \{0,1\}$. Hence $\Omega = \{0,1\}^5$. Since the coin is fair, each outcome is equally likely: for $\omega \in \Omega$ we have $P\{\omega\} = 2^{-5}$.

Similarly, if we fix a positive integer n then the probability space for flipping a fair coin n times is $\Omega = \{0,1\}^n$, and each outcome ω has probability 2^{-n}.

For an example of a concrete calculation, what is the probability that in ten coin flips, the second, fourth and seventh flips are tails? (No restrictions on the other flips.) Call this event B. In set notation,

$$B = \{\omega = (s_1, \dots, s_{10}) : s_2 = s_4 = s_7 = 1, s_i \in \{0,1\} \text{ for } i \notin \{2,4,7\}\}.$$

Since all outcomes are equally likely,

$$P(B) = \frac{\#B}{\#\Omega} = \frac{2^7}{2^{10}} = \frac{1}{8}.$$

$\#B = 2^7$ because seven of the coin flips are not constrained. △

In general, one might be able to set up several different probability spaces for a given problem. We demonstrate this with an example that shows up often in introductory probability.

Example 1.7. Consider a lottery where four distinct numbers are drawn randomly from the set $\{1, 2, 3, \dots, 30\}$. For a concrete physical experiment, imagine a rotating drum with thirty balls labeled 1 through 30, and some mechanism for drawing four balls randomly from the drum. What is the probability that the four numbers consist of three even and one odd number?

This example differs from the die rolls and coin flips in two respects.

- First, while the die draws a number repeatedly from the same set $\{1, \dots, 6\}$, now each draw changes the set of remaining options. The former is *sampling with replacement* and the latter *sampling without replacement.*
- Second, in the die and coin examples we drew ordered samples. Presently we ask only about the numbers included in the sample and not about their order of appearance. This type of question can be solved in two ways, by imagining an ordered or an unordered sample. We present both.

Solution by an ordered sample. The sample space Ω is the set of ordered 4-tuples of distinct numbers from $\{1,\dots,30\}$:

$$\Omega = \{\omega = (s_1,\dots,s_4) : \text{each } s_i \in \{1,\dots,30\} \text{ and } s_1,\dots,s_4 \text{ are distinct}\}$$

with cardinality $\#\Omega = 30 \cdot 29 \cdot 28 \cdot 27 = 657{,}720$. The fundamental assumption is again that all outcomes are equally likely, and so $P\{\omega\} = \frac{1}{657{,}720}$ for each $\omega \in \Omega$. Examples of sample points include $\omega = (23, 6, 16, 12)$ and $\omega' = (28, 29, 2, 7)$. The 4-tuple $(7, 11, 20, 11)$ is *not* an element of Ω because it repeats the entry 11.

Now count the number of 4-tuples $\omega \in \Omega$ that consist of three evens and one odd. There are 4 spots in ω for the odd number and 15 choices for this odd number. The remaining three spots in ω are then filled from left to right by even numbers, and the number of choices is $15 \cdot 14 \cdot 13$. Thus

$$P(\text{three evens and one odd}) = \frac{4 \cdot 15 \cdot 15 \cdot 14 \cdot 13}{30 \cdot 29 \cdot 28 \cdot 27} = \frac{65}{261} \approx 0.249.$$

Solution by an unordered sample. The sample space Ω is now the collection of 4-element subsets of the set $\{1,\dots,30\}$:

$$\Omega = \{\omega \subset \{1,\dots,30\} : \#\omega = 4\}$$

with cardinality $\#\Omega = \binom{30}{4} = 27{,}405$. With equally likely outcomes $P\{\omega\} = \frac{1}{27{,}405}$ for each $\omega \in \Omega$. Examples of sample points include $\omega = \{23, 6, 16, 12\}$ and $\omega' = \{28, 29, 2, 7\}$. The set $\{23, 7, 18\}$ has only three elements and hence cannot be an element of Ω. The object $\{7, 11, 20, 11\}$ is not a sensible 4-element set because of the repetition.[1]

The number of 4-element subsets of $\{1,\dots,30\}$ with three even and one odd number is $\binom{15}{3} \cdot \binom{15}{1} = 455 \cdot 15 = 6825$, obtained by first choosing a set of 3 out of 15 even numbers and then 1 number out of 15 odd numbers. Thus

$$P(\text{three evens and one odd}) = \frac{\binom{15}{3} \cdot \binom{15}{1}}{\binom{30}{4}} = \frac{65}{261} \approx 0.249.$$

Why do we get the same answer? It might seem strange that the two different approaches gave the same answer. One way to explain this is by noting that the two probability spaces are *consistent* in the following sense. A specific outcome $\{s_1, s_2, s_3, s_4\}$ in the probability space of the unordered samples can be considered an event

$$A_{s_1,s_2,s_3,s_4} = \{\omega = (a_1, a_2, a_3, a_4) : (a_1, a_2, a_3, a_4) \text{ is a permutation of } (s_1, s_2, s_3, s_4)\}.$$

in the probability space of the ordered samples. The probability of $\{s_1, s_2, s_3, s_4\}$ in the probability space of the unordered samples is $\frac{1}{\binom{30}{4}} = \frac{1}{27405}$ as this is one of the $\binom{30}{4}$ equally likely outcomes. The probability of A_{s_1,s_2,s_3,s_4} in the probability space of the unordered samples can be computed by first counting the number of elements in the event: $4! = 24$, and dividing it with the total number of outcomes $30 \cdot 29 \cdot 28 \cdot 27$ in the sample space. This gives $\frac{4!}{30 \cdot 29 \cdot 28 \cdot 27} = \frac{1}{\binom{30}{4}} = \frac{1}{27405}$, which is the same probability that we got in the probability space of the unordered samples.

[1] Collections such as $\{7, 11, 20, 11\}$ with repeated entries are called *multisets.*

This shows that if we consider an event that does not depend on the order of the sample (i.e. we can consider it in both probability spaces) then the probability of that event will be the same in both cases. Hence we can use either approaches to compute the probability.

Additional notes. Here is a third probability space that mixes the features of the first and the second example. Let

$$\Omega = \{\omega = (s_1, \dots, s_4) : \text{each } s_i \in \{1, \dots, 30\} \text{ and } s_1, \dots, s_4 \text{ are distinct}\},$$

just as in the example of the ordered sample, but set $\mathcal{F}$ to be the collection of subsets A of Ω that have the property that if $(s_1, s_2, s_3, s_4) \in A$ then $(s_{\pi_1}, s_{\pi_2}, s_{\pi_3}, s_{\pi_4}) \in A$ as well, for any permutation $(\pi_1, \pi_2, \pi_3, \pi_4)$ of $(1, 2, 3, 4)$. Thus, if $(1, 4, 7, 9) \in A$ then A must contain all $4! = 24$ different orderings of $1, 4, 7, 9$. (E.g. $(9, 1, 4, 7)$, $(1, 4, 7, 9)$, etc.) In other words: we mimic the unordered samples of the second example in the sample space of ordered samples. We use the same probability measure as in the first example, $P(A) = \frac{\#A}{\#\Omega}$. This is not a natural choice for a probability space, the only reason to mention it is because we can give an example of a function $\Omega \to \mathbb{R}$ which is **not** a random variable. Consider the function $X(s_1, s_2, s_3, s_4) = s_1$. This is a function on Ω, but it is not a random variable in our probability space. Indeed, if we consider

$$A = \{\omega : X(\omega) \le 1\} = \{\omega : X(\omega) = 1\}$$

then this is a subset of Ω which is not an element of $\mathcal{F}$: for example $(1, 2, 3, 4) \in A$, but the rearranged version $(2, 1, 3, 4)$ is not in A. △

These examples were finite. Next a countably infinite sample space.

Example 1.8. Roll a fair die until we see the first six. Record the number of rolls needed as the outcome of the experiment, and if six never comes up, record the outcome as ∞ (infinity). Thus

$$\Omega = \{\infty, 1, 2, 3, \dots\} = \mathbb{Z}_{>0} \cup \{\infty\}.$$

We derive the probability measure by comparing it to Example 1.5. When an event only depends on the outcome of the first k rolls then we can identify its probability by pretending that our experiment only includes k rolls, and then computing the probability of the event in the probability space of k die rolls as the ratio of the number of favorable outcomes over the total number of outcomes.

The event $\{k\}$ describes the outcome in our experiment where the first six came up in the kth roll. In other words, $\{k\}$ is the same as the event

$$\{\text{rolls } 1, \dots, k-1 \text{ are not six, and the } k\text{th roll is a six}\}.$$

We can compute the probability of this event using the previous example as

$$\begin{aligned} P\{k\} &= P\{\text{rolls } 1, \dots, k-1 \text{ are not six, and the } k\text{th roll is a six}\} \\ &= \frac{\#(\{1,2,3,4,5\}^{k-1} \times \{6\})}{\#(\{1,2,3,4,5,6\}^k)} = \frac{5^{k-1} \cdot 1}{6^k} = \left(\tfrac{5}{6}\right)^{k-1} \tfrac{1}{6}. \end{aligned}$$

The remaining value $P\{\infty\}$ can be derived from the axioms:

$$1 = P(\Omega) = P\{\infty, 1, 2, 3, \dots\} = P\{\infty\} + \sum_{k=1}^{\infty} P\{k\} = P\{\infty\} + \sum_{k=1}^{\infty} \left(\tfrac{5}{6}\right)^{k-1} \tfrac{1}{6}$$
$$= P\{\infty\} + 1,$$

which implies that $P\{\infty\} = 0$.

Let us calculate the probability that it takes more than m rolls of the die to get a six, where m is some fixed positive integer. This event is the union of the pairwise disjoint events $\{k\}$ for $k > m$.

$$P\{\text{more than } m \text{ rolls needed to get a six}\} = P\{m+1, m+2, m+3, \dots\}$$
$$= \sum_{k=m+1}^{\infty} P\{k\} = \sum_{k=m+1}^{\infty} \left(\tfrac{5}{6}\right)^{k-1} \tfrac{1}{6} = \left(\tfrac{5}{6}\right)^{m}.$$

Note the equivalent way to answer this question:

$$P\{\text{more than } m \text{ rolls needed to get a six}\} = P\{\text{first } m \text{ rolls are not six}\}$$
$$= \frac{5^m}{6^m} = \left(\tfrac{5}{6}\right)^{m}.$$

Note that our probability space is fairly 'sparse', as we only recorded the outcomes (the number of rolls needed for the first six). In the actual computation we looked at the richer picture: the outcomes of all the die rolls. In Example 1.12 we will see how we can properly set up a probability space that contains all that information. △

Finite and countably infinite sample spaces are called *discrete*. As the examples above illustrate, in a discrete sample space a probability measure can be defined by giving the probabilities of individual sample points. The countable additivity axiom (1.1) then determines the probabilities of all the other events through

$$P(A) = \sum_{\omega :\, \omega \in A} P\{\omega\}.$$

In uncountable sample spaces it often happens that individual sample points have probability zero. Probabilities of events cannot then be derived from the probabilities of sample points, and some other ideas are needed.

Example 1.9. Pick a real number uniformly at random from the closed unit interval $[0, 1]$. We take *uniformly at random* to mean that the chosen number is equally likely to lie anywhere in $[0, 1]$. The sample space $\Omega = [0, 1]$.

Suppose some real number $x \in [0, 1]$ has positive probability $c = P\{x\} > 0$. By the uniformity assumption, each number $y \in [0, 1]$ must then have the same probability $c = P\{y\}$. If A is a set with k elements then additivity forces $P(A) = kc$. If we take $k > 1/c$ then $P(A) > 1$. This is not allowed by the axioms. This contradiction forces us to conclude that the probability of each individual point is zero:

$$P\{x\} = 0 \quad \text{for each} \quad x \in [0, 1]. \tag{1.6}$$

Thus it is clear that we cannot base the definition of P on the probabilities of individual points. Instead, we base the definition of P on *length*. Since the length of $\Omega = [0,1]$ is 1, we get a sensible probability by stipulating that for any interval I contained in $[0,1]$, $P(I) =$ length of I. Whether the endpoints are included or excluded in the interval makes no difference. So for example

$$P([a,b]) = P((a,b)) = b-a \quad \text{for } 0 \le a < b \le 1. \tag{1.7}$$

Note the following issue related to the countable additivity axiom (1.1). The interval $[0,1]$ is a union of all the pairwise disjoint singleton sets $\{x\}$ over $x \in [0,1]$, that is, $[0,1] = \bigcup_{x\in[0,1]}\{x\}$. But we cannot claim that $P([0,1]) = \sum_{x\in[0,1]} P\{x\}$ because $P([0,1]) = 1$ while $P\{x\} = 0$ for all x. This is not a violation of axiom (1.1) because the axiom applies only to sequences of sets while $[0,1] = \bigcup_{x\in[0,1]}\{x\}$ is a union of uncountably many sets.

Intervals do not form a σ-algebra. In this example we cannot give a complete description of a probability space $(\Omega, \mathcal{F}, P)$, in contrast with the earlier examples of discrete probability spaces. There does exist a σ-algebra $\mathcal{F}$ on $[0,1]$ that contains all intervals, and a probability measure P on $\mathcal{F}$ that satisfies (1.7). This σ-algebra is the *Borel σ-algebra*. Knowing this is good enough for us, and allows us to treat all the examples we want. ♣ △

Remark 1.10. The σ-algebra that serves all our needs on the real line $\mathbb{R}$, in any Euclidean space $\mathbb{R}^d$, and on all their subsets, is called the *Borel σ-algebra* and its members are called *Borel sets*. In general, the Borel σ-algebra on any topological space is the smallest σ-algebra that contains all the open subsets of that space. A well-defined notion of length, area and volume exist for all Borel sets in $\mathbb{R}^d$. These issues are addressed in measure theory and discussed further in Section 1.5. ♣ △

The notion of a uniform random point chosen from an interval extends readily to higher dimension. Here is an example in two dimensions.

Example 1.11. Fix a positive real r_0. Pick a uniform random point from the disk of radius r_0 centered at the origin. An example of a concrete experiment could be a dart randomly thrown on a circular dart board of radius r_0 inches. The sample space is the disk itself: $\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \le r_0^2\}$.

Let A be the event that the distance from the center to the randomly chosen point is at most $r_0/5$. In two dimensions it makes sense to represent probability of a uniformly random point in terms of area. The total area of Ω is πr_0^2 which is not necessarily 1, so we take the ratio of the area of the event to the total:

$$P(A) = \frac{\text{area of } A}{\text{area of } \Omega} = \frac{\pi(r_0/5)^2}{\pi r_0^2} = \frac{1}{25}.$$

(Note that the distance of the dart from the center is a random variable: $R(x,y) = \sqrt{x^2+y^2}$.) △

The shortcomings of Example 1.11 are the same as those of Example 1.9. We cannot give a complete description of the probability space without measure-theoretic tools. But we can still perform computations. The mathematical facts left out are that there does exist a probability measure P on the Borel σ-algebra of the disk Ω such that

$$P(A) = \frac{\text{area of } A}{\text{area of } \Omega}$$

holds for each Borel set. These two examples can further be extended to define the probability space for a uniformly chosen point from a set $A \subset \mathbb{R}^n$ which has a finite n-dimensional volume.

One of the main goals of probability theory is to describe what happens to random processes over the long term. The next example describes the model required for studying the long-term behavior of rolls of a fair die.

Example 1.12. Imagine the experiment of rolling a fair die infinitely many times. This is not a realistic physical experiment, but it is a perfectly meaningful *thought experiment*. Mathematically the generalization of the sample space (1.5) of N rolls to infinitely many rolls is straightforward: N-tuples of die rolls are replaced with infinite sequences of die rolls:

$$\Omega = \{\omega = (s_k)_{1 \le k < \infty} : s_k \in \{1,2,3,4,5,6\} \ \forall k \in \mathbb{Z}_{>0}\}. \tag{1.8}$$

This space can be denoted by $\{1,2,3,4,5,6\}^\infty$ to indicate that it is an infinite Cartesian product, or by $\{1,2,3,4,5,6\}^{\mathbb{Z}_{>0}}$. This latter is notation for the space of functions from $\mathbb{Z}_{>0}$ into the set $\{1,2,3,4,5,6\}$, which is exactly what $\{1,2,3,4,5,6\}$-valued sequences are.

To describe the probability measure P on Ω of (1.8), we proceed as in the previous uncountable Examples 1.9 and 1.11, by writing down probabilities for events that we can understand. If an event only depends on the outcome of the first n die rolls then we compute its probability by pretending that the experiment only contains these n die rolls. (This is the same idea that we used in Example 1.8.) Then we can compute the probability in question by computing the ratio of the number of favorable outcomes over the total number of outcomes in n die rolls.

For example, for any finite sequence $\mathbf{t} = (t_1, \dots, t_n) \in \{1, \dots, 6\}^n$, we can define the event

$$A_{n,\mathbf{t}} = \{\omega = (s_k)_{1 \le k < \infty} \in \Omega : (s_1, \dots, s_n) = (t_1, \dots, t_n)\} \tag{1.9}$$

that prescribes the values of the first n rolls as $t_1, t_2, \dots, t_n$ in this order. We compute the probability of this event by pretending that our experiment only contains n die rolls, by Example 1.5 we get $P(A_{n,\mathbf{t}}) = 6^{-n}$.

To illustrate, suppose $n = 3$ and $\mathbf{t} = (5,1,4)$. Then

$$A_{3,(5,1,4)} = \{\omega = (s_k)_{1 \le k < \infty} \in \Omega : (s_1, s_2, s_3) = (5,1,4)\}$$

is the event that the first three die rolls come out as 5, 1, and 4. Examples of particular sample points in the set $A_{3,(5,1,4)}$ are $\omega' = (5,1,4,5,6,\dots)$ and $\omega'' = (5,1,4,2,1,1,6,\dots)$ where the subsequent entries $\dots$ can be any numbers from $\{1,\dots,6\}$. The probability is $P(A_{3,(5,1,4)}) = 6^{-3}$.

The mathematical fact that we accept without proof is the analogue of what was stated in Example 1.9. Namely, there exists a σ-algebra $\mathcal{F}$ on Ω that contains all events of the type $A_{n,\mathbf{t}}$ (for all $n \in \mathbb{Z}_{>0}$ and all $\mathbf{t} \in \{1,\dots,6\}^n$) and a probability measure P on $\mathcal{F}$ such that $P(A_{n,\mathbf{t}}) = 6^{-n}$. This σ-algebra is known as the *product σ-algebra*. ♣

Here is an example of a concrete calculation. What is the probability that the first five die rolls are all either 1 or 2, and the next five die rolls are neither 1 nor

2? Call this event B. B involves the first ten die rolls. We can express B as a the disjoint union

$$B = \bigcup_{\mathbf{t}\in\mathcal{H}} A_{10,\mathbf{t}}$$

where $\mathcal{H}$ is the set of 10-tuples $\mathbf{t}$ whose first five coordinates are 1 or 2, and whose last five coordinates are neither 1 nor 2:

$$\begin{aligned}\mathcal{H} = \{\mathbf{t} \in \{1,\dots,6\}^{10} : t_i \in \{1,2\} \text{ for } i = 1,\dots,5,\\ \text{and } t_j \in \{3,4,5,6\} \text{ for } j = 6,\dots,10\}.\end{aligned}$$

The multiplication principle of counting gives $\#\mathcal{H} = 2^5 \cdot 4^5$. Since each event $A_{10,\mathbf{t}}$ has probability 6^{-10}, we get

$$\begin{aligned}&P(\text{first 5 die rolls are 1 or 2, next 5 die rolls are neither 1 nor 2})\\ &= P(B) = \sum_{\mathbf{t}\in\mathcal{H}} P(A_{10,\mathbf{t}}) = 2^5 \cdot 4^5 \cdot 6^{-10} = \tfrac{2^5}{3^{10}} = \tfrac{32}{59049}.\end{aligned}$$

Note that the computation boiled down to counting 'favorable' 10-tuples and dividing this number with 6^{10}. This is because the probability of the event B in our probability space is the same as the probability of the corresponding event in the probability space of 10 die rolls.

Note that we can use the probability space of infinitely many die rolls to model the experiment described in Example 1.8. We have

$$\{k\} = \{w = (s_i)_{1\le i<\infty} \in \Omega : s_k = 6, s_1, \dots, s_{k-1} \neq 6\}$$

for each positive integer k, and

$$\{\infty\} = \{w = (s_i)_{1\le i<\infty} \in \Omega : s_i \neq 6\}.$$

△

Example 1.13. The probability space for infinitely many fair coin flips can be constructed using the ideas in Examples 1.6 and 1.12.

The sample space is the space of $\{0,1\}$-valued sequences:

$$\Omega = \{0,1\}^{\mathbb{Z}_{>0}} = \{\omega = (s_k)_{1\le k<\infty} : s_k \in \{0,1\}\ \forall k \in \mathbb{Z}_{>0}\}. \tag{1.10}$$

As in Example 1.12, for a positive integer n and an n-tuple $\mathbf{t} = (t_1, \dots, t_n) \in \{0,1\}^n$, define the event $A_{n,\mathbf{t}}$ as the set of sequences ω whose first n entries follow $\mathbf{t}$:

$$A_{n,\mathbf{t}} = \{\omega = (s_k)_{1\le k<\infty} \in \Omega : (s_1, \dots, s_n) = (t_1, \dots, t_n)\}. \tag{1.11}$$

To be consistent with Example 1.6, we must stipulate that $P(A_{n,\mathbf{t}}) = 2^{-n}$ for all $n \in \mathbb{Z}_{>0}$ and $\mathbf{t} \in \{0,1\}^n$.

For example, the event that the first five flips are all tails would be

$$A_{5,(1,1,1,1,1)} = \{\omega = (s_k)_{1\le k<\infty} \in \Omega : (s_1, \dots, s_5) = (1,1,1,1,1)\}$$

and its probability $P(A_{5,(1,1,1,1,1)}) = 2^{-5}$.

We take for granted that this is a sensible description of a probability space, even though we do not fully describe a σ-algebra $\mathcal{F}$ on Ω and a probability measure P on $\mathcal{F}$. △

Remark 1.14. Being familiar with the precise definition of a probability space is important for everybody who wants work with problems related to probability. However, in practice we will mostly deal with the first and last component of the triple $(\Omega, \mathcal{F}, P)$: the sample space Ω and the probability measure P.

In most of our applications in this textbook the probability space we will work with falls into one of the following two cases:

1. The probability space is discrete and all singletons are events. In this case $\mathcal{F}$ is the set of all subsets of Ω, and the probability measure can be described with the probabilities of the singletons.

2. The probability space is a version of the ones seen in Examples 1.9, 1.11, 1.12, and 1.13. In this case we identify 'nice' subsets of Ω for which we have clean formulas for the probability (e.g. subintervals in Example 1.9, sets that only depend on finitely many coordinates in Examples 1.12 and 1.13). We do not have an explicit description of all events and the corresponding probabilities, but more advanced measure theoretic techniques can show that one can define $\mathcal{F}$ and P in a way that $\mathcal{F}$ is the set of all sets that can be obtained from the nice sets (via the appropriate set operations), and P gives the appropriate probabilities for the 'nice' sets. △

1.2. Properties of probability measures

In this section we discuss several consequences of the probability axioms that are both theoretically important and useful for practical computations. An immediately useful property is the additivity of probability. We have seen this before, but it is important enough that we restate it:

Fact 1.15. Suppose $A_1, A_2, \dots$ are events in the same probability space. If for some (finite or infinite) k the events A_j for $1 \le j \le k$ are mutually exclusive (in other words: pairwise disjoint) then

$$P(\cup_{j=1}^{k} A_j) = \sum_{j=1}^{k} P(A_j). \tag{1.12}$$

This statement holds for finite k or $k = \infty$, but it is important that the events in question must be disjoint.

Calculation of the probability of a complicated event typically requires decomposing the event into smaller mutually exclusive pieces whose probabilities are easier to evaluate. The next example illustrates.

Example 1.16. Roll a fair die until the first even number appears. What is the probability that this number is a four?

The answer is of course $\frac{1}{3}$ since 2, 4 and 6 should be equally likely. But let us derive it by decomposing the event.

Let $A = \{\text{the first even number is a four}\}$ be the desired event. Since we know how to calculate probabilities of events that involve a fixed number of rolls, let us decompose A according to the number of rolls needed to see the first even number. So let

$$A_k = \{\text{the } k\text{th roll is a four, and rolls } 1, \dots, k-1 \text{ yield odd numbers}\}.$$

Then $A = \bigcup_{k \geq 1} A_k$ expresses A as a union of pairwise disjoint events A_k. By counting favorable arrangements,

$$P(A_k) = \frac{3^{k-1} \cdot 1}{6^k} = \left(\tfrac{1}{2}\right)^{k-1} \tfrac{1}{6}.$$

Then

$$P(A) = \sum_{k=1}^{\infty} P(A_k) = \sum_{k=1}^{\infty} \left(\tfrac{1}{2}\right)^{k-1} \tfrac{1}{6} = \tfrac{1}{6} \cdot \frac{1}{1 - \frac{1}{2}} = \tfrac{1}{3}.$$

For a slightly more complicated question, what is the probability that the first two even numbers we see are both fours? The desired event is

$$B = \{\text{the first two even numbers are fours}\}.$$

To decompose B into manageable pieces, define for integers $k > j \geq 1$ events

$$\begin{aligned} B_{j,k} = \{&\text{the } j\text{th and } k\text{th rolls are fours, and} \\ &\quad \text{rolls } 1, \dots, j-1, j+1, \dots, k-1 \text{ yield odd numbers}\}. \end{aligned}$$

Then

$$P(B_{j,k}) = \frac{3^{k-2} \cdot 1^2}{6^k} = \left(\tfrac{1}{2}\right)^{k-2} \tfrac{1}{36}$$

and from the disjoint union $B = \bigcup_{k > j \geq 1} B_{j,k}$ we have

$$\begin{aligned} P(B) &= \sum_{k > j \geq 1} P(B_{j,k}) = \sum_{j=1}^{\infty} \sum_{k=j+1}^{\infty} \left(\tfrac{1}{2}\right)^{k-2} \tfrac{1}{36} = \tfrac{1}{36} \sum_{j=1}^{\infty} \frac{\left(\frac{1}{2}\right)^{j-1}}{1 - \frac{1}{2}} \\ &= \tfrac{1}{18} \sum_{j=1}^{\infty} \left(\tfrac{1}{2}\right)^{j-1} = \tfrac{1}{18} \cdot \frac{1}{1 - \frac{1}{2}} = \tfrac{1}{9}. \end{aligned}$$

Explain why the answer $\frac{1}{9}$ should not be surprising. △

For any event A the events A, A^c are disjoint, and their union is equal to Ω. This means that

$$P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1,$$

which leads to the following useful identity:

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

Next we develop properties that need proof and illustrate their use. The first theorem gives an algebraic identity, the second theorem gives two inequalities, and the third theorem gives two limits.

Theorem 1.17. *Let $A_1, A_2, A_3, \dots$ be events in some probability space $(\Omega, \mathcal{F}, P)$. Then for each integer $n \geq 2$,*

$$\begin{aligned} P(A_1 \cup \cdots \cup A_n) \ &= \sum_{i=1}^{n} P(A_i) \ - \sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) \\ &\quad - \sum_{1 \le i_1 < i_2 < i_3 < i_4 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3} \cap A_{i_4}) \\ &\quad + \ \cdots \ + \ (-1)^{n+1} P(A_1 \cap \cdots \cap A_n). \\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} P(A_{i_1} \cap \cdots \cap A_{i_k}). \end{aligned} \tag{1.14}$$

This is called the inclusion-exclusion identity.

The last line of (1.14) is simply a convenient way to present all the terms together. Before we turn to the proof, a few comments. When events are not pairwise independent, additivity (1.2) cannot be applied. The inclusion-exclusion identity (1.14) tells us exactly how to account for the overcounting of overlaps. If $A_1, \dots, A_n$ are pairwise disjoint, (1.14) reduces to (1.2) because on the right-hand side of (1.14) all the intersections have probability zero.

To understand formula (1.14), note for example that the sum

$$\sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} \cap A_{i_2} \cap A_{i_3})$$

ranges over all triples of distinct indices. There are $\binom{n}{3}$ such triples. The inequality $i_1 < i_2 < i_3$ in the sum ensures that each triple is counted only once. For $n = 2$ and $n = 3$ the inclusion-exclusion identity gives:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{1.15}$$

and

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ - P(B \cap C) + P(A \cap B \cap C). \end{aligned} \tag{1.16}$$

Proof of Theorem 1.17. We do this by induction on n. The base case is $n = 2$, namely (1.15), which follows from these observations:

$$A = AB \cup AB^c, \quad B = AB \cup A^cB, \quad \text{and} \quad A \cup B = AB \cup AB^c \cup A^cB,$$

and in each case the right-hand side is a union of disjoint events. Now start with the last identity, add and subtract $P(AB)$ once, and collect terms:

$$\begin{aligned} P(A \cup B) &= P(AB) + P(AB^c) + P(A^cB) \\ &= \big(P(AB) + P(AB^c)\big) + \big(P(AB) + P(A^cB)\big) - P(AB) \\ &= P(A) + P(B) - P(AB). \end{aligned}$$

For the induction step, we assume the cases for n events and two events and prove the case for $n+1$ events.

$$\begin{aligned} P\Big(\bigcup_{k=1}^{n+1} A_k\Big) &= P\Big(\Big(\bigcup_{k=1}^{n} A_k\Big) \cup A_{n+1}\Big) \\ &= P\Big(\bigcup_{k=1}^{n} A_k\Big) + P(A_{n+1}) - P\Big(\bigcup_{k=1}^{n} (A_k \cap A_{n+1})\Big) \\ &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} P(A_{i_1}\cap\cdots\cap A_{i_k}) \;+\; P(A_{n+1}) \\ &\qquad - \sum_{\ell=1}^{n} (-1)^{\ell+1} \sum_{1\le i_1<\cdots<i_\ell\le n} P(A_{i_1}\cap\cdots\cap A_{i_\ell}\cap A_{n+1}). \end{aligned}$$

From the first sum, separate out the term $k=1$ and combine it with $P(A_{n+1})$. In the last sum, separate out the term $\ell = n$. In the remaining sum ℓ ranges from 1 to $n-1$. In this remaining sum, set $-(-1)^{\ell+1} = (-1)^{\ell+2}$, change the summation index to $k=\ell+1$ which ranges from 2 to n, and denote the index $n+1$ as $i_{\ell+1} = i_k$. Then we have

$$\begin{aligned} P\Big(\bigcup_{k=1}^{n+1} A_k\Big) \;&=\; \sum_{i=1}^{n+1} P(A_i) \;+\; \sum_{k=2}^{n} (-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} P(A_{i_1}\cap\cdots\cap A_{i_k}) \\ &\qquad + \sum_{k=2}^{n} (-1)^{k+1} \sum_{1\le i_1<\cdots<i_{k-1}<i_k=n+1} P(A_{i_1}\cap\cdots\cap A_{i_k}) \\ &\qquad + (-1)^{n+1} P(A_1\cap\cdots\cap A_{n+1}). \end{aligned}$$

Now the two middle sums combine to give all the choices of k distinct indices from $1,\dots,n+1$. The first middle sum contains the choices restricted to $\{1,\dots,n\}$, while the second middle sum contains the choices that include $n+1$. Thus the right-hand side of the formula is exactly the desired right-hand side in (1.14) for the case of $n+1$ events. This completes the proof. □

Example 1.18. What is the probability that four fair die rolls yield at least one six?

$$P(\text{at least one six}) = 1 - P(\text{no sixes}) = 1 - \left(\tfrac{5}{6}\right)^4.$$

Turn over the top card of a well-shuffled deck of 52 cards. What is the probability that the card is either a face card or a spade? *Well-shuffled* means that each card is equally likely. *Face cards* are kings, queens and jacks.

$$\begin{aligned} &P(\texttt{face card} \text{ or } \texttt{spade}) \\ &\qquad = P(\texttt{face card}) + P(\texttt{spade}) - P(\texttt{face card} \text{ and } \texttt{spade}) \\ &\qquad = \tfrac{12}{52} + \tfrac{13}{52} - \tfrac{3}{52} = \tfrac{22}{52} = \tfrac{11}{26}. \end{aligned}$$

△

The next example is a probability classic.

Example 1.19. Suppose n people arrive for a show and leave their hats in the cloakroom. The cloakroom attendant mixes up the hats completely so that each person leaves with a random hat. Let us assume that all $n!$ assignments of hats are equally likely. What is the probability that no one gets their own hat? How does this probability behave as $n \to \infty$?

Define the events

$$A_i = \{\text{person } i \text{ gets his/her own hat}\}, \quad 1 \le i \le n.$$

The probability we want is

$$P\Big(\bigcap_{i=1}^{n} A_i^c\Big) = 1 - P\Big(\bigcup_{i=1}^{n} A_i\Big), \tag{1.17}$$

where we used de Morgan's law. We compute $P(A_1 \cup \cdots \cup A_n)$ with the inclusion-exclusion formula (1.14). With n fixed, let $1 \le k \le n$, take k distinct indices $i_1 < i_2 < \cdots < i_k$ in the range $\{1, \dots, n\}$, and evaluate

$$\begin{aligned} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) &= P\{\text{individuals } i_1, i_2, \dots, i_k \text{ get their own hats}\} \\ &= \frac{(n-k)!}{n!}. \end{aligned} \tag{1.18}$$

The formula comes from counting favorable arrangements: if k hats are assigned, there are $(n-k)!$ ways to distribute the remaining $n-k$ hats. (The event $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$ places no restrictions on these other $n-k$ hats.) Thus

$$\sum_{i_1<i_2<\cdots<i_k} P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \binom{n}{k}\frac{(n-k)!}{n!} = \frac{1}{k!},$$

since there are $\binom{n}{k}$ terms in the sum. From (1.14)

$$P\Big(\bigcup_{i=1}^{n} A_i\Big) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots + (-1)^{n+1}\frac{1}{n!}$$

and then by (1.17)

$$P\Big(\bigcap_{i=1}^{n} A_i^c\Big) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \cdots + (-1)^n\frac{1}{n!} = \sum_{k=0}^{n}\frac{(-1)^k}{k!}. \tag{1.19}$$

Thus the limit as $n \to \infty$ is immediate:

$$\lim_{n\to\infty} P(\text{no person among } n \text{ people gets the correct hat}) = \sum_{k=0}^{\infty}\frac{(-1)^k}{k!} = e^{-1}.$$

Mathematically this example is about the number of fixed points of a *random permutation*. If f is a permutation of the set $\{1, \dots, n\}$ (that is, a bijective function from $\{1, \dots, n\}$ onto $\{1, \dots, n\}$), then the *fixed points* of f are those elements x that satisfy $f(x) = x$. In the example, $f(i) = j$ if person i receives the hat of person j. A fixed point $f(i) = i$ means that person i receives the correct hat. The result says that as $n \to \infty$, the probability that a uniformly random permutation on n elements has no fixed points converges to e^{-1}. △

Theorem 1.20. *Let A, B, $A_1, A_2, A_3, \dots$ be events in some probability space $(\Omega, \mathcal{F}, P)$.*

(i) *Monotonicity: if $A \subset B$ then $P(A) \le P(B)$.*

(ii) *Countable subadditivity: for any sequence of events $\{A_k\}$,*

$$P\Big(\bigcup_{k=1}^{\infty} A_k\Big) \le \sum_{k=1}^{\infty} P(A_k). \tag{1.20}$$

Countable subadditivity (1.20) generalizes the countable additivity axiom (1.1) in a natural way. Its truth should be fairly obvious because the union $\bigcup_{k=1}^{\infty} A_k$ can have overlaps whose probabilities are then counted several times over in the sum $\sum_{i=1}^{\infty} P(A_k)$. By taking $A_k = \varnothing$ for all $k > n$ we get a finite version of subadditivity:

$$P(A_1 \cup \cdots \cup A_n) \le P(A_1) + \cdots + P(A_n), \tag{1.21}$$

valid for all events $A_1, \dots, A_n$.

Proof of Theorem 1.20. (i) A and $B \setminus A$ are disjoint and $A \cup (B \setminus A) = B$. Hence

$$P(B) = P(A) + P(B \setminus A) \ge P(A). \tag{1.22}$$

(ii) Let $A = \bigcup_{k=1}^{\infty} A_k$. We define a new sequence of events that enables us to write A as a union of pairwise disjoint events. Let

$$\begin{aligned}
B_1 &= A_1,\\
B_2 &= A_2 \setminus A_1 = A_1^c \cap A_2,\\
B_3 &= A_3 \setminus (A_1 \cup A_2) = A_1^c \cap A_2^c \cap A_3, \text{ and so on,}\\
B_k &= A_k \setminus (A_1 \cup \cdots \cup A_{k-1}) = A_1^c \cap \cdots \cap A_{k-1}^c \cap A_k
\end{aligned}$$

for all indices k. In other words, B_k consists of those sample points in A_k that have not already appeared in one of the earlier events $A_1, \dots, A_{k-1}$. Events B_k are pairwise disjoint because if $k < \ell$ then $B_k \subset A_k$ while $B_\ell \subset A_k^c$.

To prove that $A = \bigcup_{k=1}^{\infty} B_k$, note first that since $B_k \subset A_k$ for each k,

$$\bigcup_{k=1}^{\infty} B_k \subset \bigcup_{k=1}^{\infty} A_k = A.$$

To show the opposite inclusion, let $\omega \in A$. By definition of the union $A = \bigcup_{k=1}^{\infty} A_k$, ω lies in one of the events A_k, and hence we can pick m as the *first* index such that $\omega \in A_m$. This definition of m means that ω does not lie in any of the events $A_1, \dots, A_{m-1}$. Hence

$$\omega \in A_m \setminus (A_1 \cup \cdots \cup A_{m-1}) = B_m.$$

Thus every $\omega \in A$ lies in some B_k, which says that $A \subset \bigcup_{k=1}^{\infty} B_k$.

Since $P(B_k) \le P(A_k)$ for each k, the countable additivity axiom gives

$$P(A) = \sum_{k=1}^{\infty} P(B_k) \le \sum_{k=1}^{\infty} P(A_k).$$

□

Example 1.21. Recall the probability space for infinitely many rolls of a fair die from Example 1.12. We show that individual sample points have probability zero, namely, that $P\{\omega\} = 0$ for each $\omega \in \Omega$.

Fix some particular $\widetilde{\omega} \in \Omega$. By definition, $\widetilde{\omega}$ is an infinite sequence of integers from the set $\{1, \dots, 6\}$. Let $t_1, t_2, t_3, \dots$ be the entries of $\widetilde{\omega}$, so that $\widetilde{\omega} = (t_k)_{k \in \mathbb{Z}_{>0}}$. Let

$$A_n = \{\omega = (s_k)_{k \in \mathbb{Z}_{>0}} \in \Omega : (s_1, \dots, s_n) = (t_1, \dots, t_n)\}$$

be the event that the first n rolls match those of $\widetilde{\omega}$. (In more mathematical terms: A_n is the set of sequences ω in Ω whose first n entries match those of $\widetilde{\omega}$.) Then, as we have calculated several times already, $P(A_n) = (\frac{1}{6})^n$. From $\{\widetilde{\omega}\} \subset A_n$ we have $P\{\widetilde{\omega}\} \le P(A_n) = \left(\frac{1}{6}\right)^n$, and this inequality is valid for all n. Since $\left(\frac{1}{6}\right)^n \to 0$ as $n \to \infty$, in the limit $P\{\widetilde{\omega}\} \le 0$, which implies $P\{\widetilde{\omega}\} = 0$. △

Example 1.22. Consider again the probability space for infinitely many rolls of a fair die. We have seen that with probability one there will be a six at some point. We now consider a generalization of this: let $(a_1, a_2, \dots, a_k)$ be a finite sequence of possible die rolls (so that $1 \le a_i \le 6$ for each $1 \le i \le k$). Let A be the event that in our infinite sequence of die rolls we see exactly the sequence $(a_1, a_2, \dots, a_k)$ appearing at some point:

$$A = \{\omega = (t_1, t_2, \dots) : \omega \in \Omega, \exists n \ge 1 \text{ so that } t_n = a_1, t_{n+1} = a_2, \dots, t_{n+k-1} = a_k\}.$$

(With $k = 1$ and $a_1 = 6$ this is just the event that we see a six at some point.) We will prove that $P(A) = 1$.

We do this by showing that $P(A^c) = 0$. For an $\ell \ge 1$ let B_ℓ denote the event that the sequence $(a_1, a_2, \dots, a_k)$ does not show up in any of the non-overlapping blocks $(t_1, \dots, t_k)$, $(t_{k+1}, \dots, t_{2k})$, $\dots$, $(t_{(\ell-1)k+1}, \dots, t_{\ell k})$. Because A^c is the event that we will never see the sequence $(a_1, a_2, \dots, a_k)$, we have

$$P(A^c) \le P(B_\ell)$$

for each $\ell \ge 1$. The event B_ℓ depends on the outcome of the first $k \cdot \ell$ die rolls, we can compute its probability just by counting. Since we want to avoid in the appearance of the sequence $(a_1, a_2, \dots, a_k)$ in any block $(t_{ik+1}, \dots, t_{(i+1)k})$ for $i = 0, 1, \dots, \ell - 1$, in each of these blocks we can choose the outcomes $6^k - 1$ different ways. Hence there $(6^k - 1)^\ell$ favorable outcomes out of the $6^{k \cdot \ell}$ possible outcomes, which gives

$$P(B_\ell) = \frac{(6^k - 1)^\ell}{6^{k \cdot \ell}} = \left(1 - 6^{-k}\right)^\ell .$$

This yields the bound

$$P(A^c) \le \left(1 - 6^{-k}\right)^\ell .$$

As $\ell \to \infty$ our upper bound goes to zero, hence $P(A^c)$ must be equal to zero, and $P(A) = 1$.

The result that we just proved appears in the popular culture as the monkey and the typewriter problem. Imagine that a monkey sits in front of a typewriter (or a laptop), and randomly starts typing without ever stopping. Then with probability one we will see your favorite novel appearing at some point as the result of the monkey's typing.

Assuming that the various keys are equally likely to be hit for each keystroke, this is just a version of the infinitely many die roll experiment, except instead of the numbers $1, 2, \dots, 6$ we get various letters (and other symbols) in each step. By our result, for any given sequence of letters/symbols (which could be for example the text of the first Harry Potter book) with probability one that sequence will appear at some point in the infinite text typed by the monkey. △

Corollary 1.23. *Let $\{A_k\}$ be a sequence of events on $(\Omega, \mathcal{F}, P)$.*

(i) *If $P(A_k) = 0$ for all k, then $P(\bigcup_k A_k) = 0$.*

(ii) *If $P(A_k) = 1$ for all k, then $P(\bigcap_k A_k) = 1$.*

Proof. (i) By (1.20),

$$P\Big(\bigcup_k A_k\Big) \le \sum_k P(A_k) = 0.$$

(ii) By de Morgan's law and then part (i) applied to $\{A_k^c\}$,

$$P\Big(\bigcap_k A_k\Big) = 1 - P\Big(\Big(\bigcap_k A_k\Big)^c\Big) = 1 - P\Big(\bigcup_k A_k^c\Big) = 1 - 0 = 1.$$

□

Example 1.24. In Example 1.9, what is the probability that the chosen number is rational? Since the rationals are countable, we can express the set $\mathbb{Q} \cap [0, 1]$ of rationals in $[0, 1]$ as a sequence: $\mathbb{Q} \cap [0, 1] = \{q_1, q_2, q_3 \dots\}$. Hence

$$P(\mathbb{Q} \cap [0, 1]) = \sum_{k=1}^{\infty} P\{q_k\} = \sum_{k=1}^{\infty} 0 = 0.$$

△

As the last item of this section, we look at how probabilities of events converge. Suppose $\{A_k\}_{k\in\mathbb{Z}_{>0}}$, $\{B_k\}_{k\in\mathbb{Z}_{>0}}$, A, and B are events in a probability space $(\Omega, \mathcal{F}, P)$. We say that A_k *increases up to* A and use the notation

$$A_k \nearrow A \tag{1.23}$$

if the events A_k are nested nondecreasing, which means that $A_1 \subset A_2 \subset A_3 \subset \cdots \subset A_k \subset \cdots$, and $A = \bigcup_k A_k$. Figure 1 illustrates.

Analogously, we say that B_k *decreases down to* B and use the notation

$$B_k \searrow B \tag{1.24}$$

if the events B_k are nested nonincreasing, which means that $B_1 \supset B_2 \supset B_3 \supset \cdots \supset B_k \supset \cdots$, and $B = \bigcap_k B_k$.

Theorem 1.25. *If $A_k \nearrow A$ or $A_k \searrow A$, then the probabilities converge: $\lim_{k\to\infty} P(A_k) = P(A)$.*

Proof. Suppose first $A_k \nearrow A$. Let $A_0 = \varnothing$. Then we can write

$$A = \bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty} (A_k \setminus A_{k-1}).$$

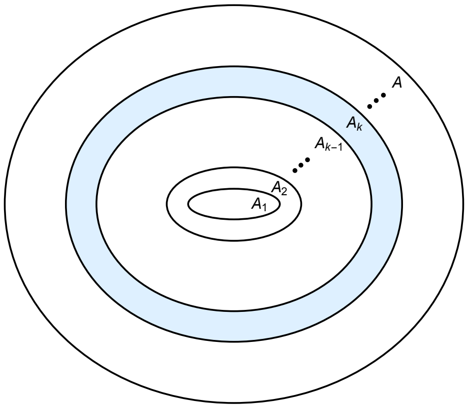

Figure 1. Illustration of $A_k \nearrow A$. The shaded portion is $A_k \setminus A_{k-1}$.

The events $A_k \setminus A_{k-1}$ are pairwise disjoint. (The details here are left to Exercise 1.13.) Then, using axiom (1.1) and the definition of the value of a series as the limit of partial sums,

$$P(A) = \sum_{k=1}^{\infty} P(A_k \setminus A_{k-1}) = \lim_{n\to\infty} \sum_{k=1}^{n} P(A_k \setminus A_{k-1})$$

$$= \lim_{n\to\infty} \sum_{k=1}^{n} \big(P(A_k) - P(A_{k-1})\big) = \lim_{n\to\infty} \big(P(A_n) - P(A_0)\big) = \lim_{n\to\infty} P(A_n).$$

Exercise 1.14 asks you to do an analogous argument for the case $A_k \searrow A$. Here we apply the already proved part to complements. $A_k \supset A_{k+1}$ implies that $A_k^c \subset A_{k+1}^c$. By de Morgan's law, $A = \bigcap_k A_k$ implies $A^c = \bigcup_k A_k^c$. Thus $A_k^c \nearrow A^c$, and the first part of the lemma gives the convergence $P(A_k^c) \to P(A^c)$. From this, $P(A_k) = 1 - P(A_k^c) \to 1 - P(A^c) = P(A)$. □

Example 1.26. Consider again Example 1.9 with sample space $\Omega = [0,1]$ given by the unit interval of the real line, and probability given by length: $P(I) = \text{length}$ of I for any interval $I \subset \Omega$. Fix a point $x \in (0,1)$ and let $A_n = (x - \frac{1}{n}, x + \frac{1}{n})$. For $n > \max\{x^{-1}, (1-x)^{-1}\}$ we have $A_n \subset \Omega$. $A_n \searrow \{x\}$ because the sets A_n are nested decreasing, each contains x, and any $y \neq x$ satisfies $y \notin A_n$ once n is so large that $\frac{1}{n} < |y - x|$. Thus

$$P\{x\} = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} \tfrac{2}{n} = 0.$$

This is a different proof of $P\{x\} = 0$ which was already argued in Example 1.9 by appeal to additivity. △

Example 1.27. We give an alternative derivation of the fact that repeated rolls of a fair die eventually produce a six. Define events

$$A = \{\text{six appears eventually}\}$$

and

$$A_k = \{\text{six appears at least once in the first } k \text{ rolls}\}.$$

Then $A_k \nearrow A$. The probability of A_k is found quickly by switching to complements:

$$P(A_k) = 1 - P\{\text{no six in the first } k \text{ rolls}\} = 1 - \frac{5^k}{6^k} = 1 - \left(\tfrac{5}{6}\right)^k.$$

By Theorem 1.25,

$$P(A) = \lim_{k\to\infty} \left(1 - \left(\tfrac{5}{6}\right)^k\right) = 1.$$

△

Example 1.28. We have a (very large) urn. At time 0 we have a red and green marble in it. At time 1 we remove one of the two marbles from the urn, and add two red marbles. We repeat that over and over again: at time k we remove one of the marbles from the urn, and add two red marbles. (So after this step we have $k+2$ marbles in the urn.) What is the probability that eventually the green marble is removed from the urn?
Let A_k be the event that the green marble is still in the urn immediately after the kth step. Then we have $A_1 \supset A_2 \supset \dots$. The set $\cap_{k=1}^{\infty} A_k$ is the event that the green marble never gets removed from the urn, so

$$P(\text{eventually the green marble is removed}) = 1 - P(\cap_{k=1}^{\infty} A_k).$$

We have

$$P(\cap_{k=1}^{\infty} A_k) = \lim_{k\to\infty} P(A_k).$$

We can compute $P(A_k)$ by counting favorable outcomes within the first k steps. We have $2 \cdot 3 \cdots (k+1) = (k+1)!$ equally likely outcomes, and in $1 \cdot 2 \cdots k = k!$ of these outcomes we always pick a red marble in each step. Hence

$$P(A_k) = \frac{k!}{(k+1)!} = \frac{1}{k+1},$$

and

$$P(\text{eventually the green marble is removed}) = 1 - \lim_{k\to\infty} \frac{1}{k+1} = 1.$$

This means that with probability one the green marble is removed at some point.

△

1.3. Conditional probability

The probability of an event A when another event B is assumed to happen is captured by the notion of conditional probability.

Definition 1.29. Let B be an event on the probability space $(\Omega, \mathcal{F}, P)$ such that $P(B) > 0$. Then for all events $A \in \mathcal{F}$ the **conditional probability of A given B** is defined as

$$P(A \,|\, B) = \frac{P(AB)}{P(B)}. \tag{1.25}$$

That this definition is natural is easy to see in an example of equally likely outcomes.

Example 1.30. Consider the experiment of three fair coin flips, with sample space

$$\Omega = \{\omega = (s_1, s_2, s_3) : \text{each } s_i \in \{0,1\}\}$$

and equally likely outcomes: $P\{\omega\} = \frac{1}{8}$. We call outcome 0 heads and outcome 1 tails.

Define the event

$$A = \{\text{exactly 2 heads}\} = \{(0,0,1), (0,1,0), (1,0,0)\}.$$

Without any information about the experiment, $P(A) = \frac{3}{8}$. Suppose that subsequently we learn the outcome of the first flip. Let

$$B = \{\text{first flip is heads}\} = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1)\}.$$

If B happens, then only four sample points are possible. Two of these lie in A, namely $(0,0,1), (0,1,0)$. Since sample points are equally likely, the new probability of A should be $\frac{2}{4} = \frac{1}{2}$. This is exactly the value that arises from the formula (1.25):

$$P(A\,|\,B) = \frac{P(AB)}{P(B)} = \frac{P\{(0,0,1),(0,1,0)\}}{P\{(0,0,0),(0,0,1),(0,1,0),(0,1,1)\}} = \frac{2/8}{4/8} = \frac{1}{2}.$$

In the opposite case that the first flip is tails,

$$P(A\,|\,B^c) = \frac{P(AB^c)}{P(B^c)} = \frac{P\{(1,0,0)\}}{P\{(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}} = \frac{1/8}{4/8} = \frac{1}{4}.$$

△

The first observation is that conditioning on an event with positive probability produces a new probability measure on the same sample space. We leave the verification of this fact as Exercise 1.18.

Theorem 1.31. *Let B be an event on the probability space $(\Omega, \mathcal{F}, P)$ such that $P(B) > 0$. Then as a function of the event A, $P(A|B)$ is a probability measure on $(\Omega, \mathcal{F})$.*

Remark 1.32. When the original probability measure P is replaced by the conditioned probability measure $P(\cdot\,|B)$, we have a new probability space $(\Omega, \mathcal{F}, P(\cdot\,|B))$. We do not replace the sample space Ω with the set B, even though the portion of Ω outside B now has probability zero: $P(B^c|B) = 0$. It turns out useful to continue to compute various conditioned quantities on the same sample space. We see this immediately below when we want to condition simultaneously on several different events, without imagining a new sample space for each conditional probability. △

Through the introduction of conditional probability we gain new formulas for calculating unconditioned probabilities. We state two in the next theorem. Part (a) of the theorem is sometimes called the *multiplication identity* and part (b) the *law of total probability.* First a definition: a finite or countably infinite collection of events $\{B_i\}$ is a *countable partition* of Ω if

$$B_i \cap B_j = \varnothing \quad \text{whenever } i \neq j \text{ and } \quad \bigcup_i B_i = \Omega. \tag{1.26}$$

In other words, the events $\{B_i\}$ are pairwise disjoint and together they make up Ω.

Theorem 1.33. *In each statement below all events are on the same probability space* $(\Omega, \mathcal{F}, P)$.

(a) *Let* A *and* B *be two events and assume* $P(B) > 0$. *Then*

$$P(AB) = P(B)P(A \,|\, B). \tag{1.27}$$

Let $A_1, \dots, A_n$ *be events and assume* $P(A_1 \cdots A_{n-1}) > 0$. *Then*

$$P(A_1 A_2 \cdots A_n) = P(A_1)\, P(A_2 \,|\, A_1)\, P(A_3 \,|\, A_1 A_2) \cdots P(A_n \,|\, A_1 \cdots A_{n-1}). \tag{1.28}$$

(b) *Let* $\{B_i\}$ *be a countable partition of* Ω *and* A *any event. Then*

$$P(A) = \sum_{i\,:\,P(B_i)>0} P(A \,|\, B_i)\, P(B_i). \tag{1.29}$$

The sum above ranges over those indices i *such that* $P(B_i) > 0$.

Proof. Part (a). Equation (1.27) is a rearrangment of (1.25). Equation (1.28) comes from algebra:

$$\begin{aligned} P(A_1 A_2 \cdots A_n) &= P(A_1) \cdot \frac{P(A_1 A_2)}{P(A_1)} \cdot \frac{P(A_1 A_2 A_3)}{P(A_2 A_1)} \cdots \frac{P(A_1 \cdots A_{n-1} A_n)}{P(A_1 \cdots A_{n-1})} \\ &= P(A_1) P(A_2 \,|\, A_1) P(A_3 \,|\, A_1 A_2) \cdots P(A_n \,|\, A_1 \cdots A_{n-1}). \end{aligned}$$

The assumption $P(A_1 \cdots A_{n-1}) > 0$ guarantees that none of the denominators vanishes.

Part (b). Given a countable partition $\{B_i\}$ of Ω, any event A satisfies

$$A = A \cap \Omega = A \cap \Bigl(\bigcup_i B_i \Bigr) = \bigcup_i AB_i. \tag{1.30}$$

This expresses A as the union of the pairwise disjoint events AB_i. To complete the proof of (1.29), use additivity of probability, then note that $P(B_i) = 0$ implies $P(AB_i) = 0$ and drop these terms from the sum, and finally apply the multiplication identity (1.27) to each probability $P(AB_i)$:

$$P(A) = \sum_i P(AB_i) = \sum_{i\,:\,P(B_i)>0} P(AB_i) = \sum_{i\,:\,P(B_i)>0} P(A|B_i)P(B_i). \qquad \square$$

Example 1.34. Formula (1.27) gives a probabilistic explanation for counting in sampling without replacement. Suppose an urn contains 5 red, 3 yellow and 2 black balls. Draw four balls one by one, without replacement. What is the probability that we see first two reds, followed by a black, and last a yellow ball?

Denote the events by R_1, R_2 (red in the first and second draws), B_3 (black in the third draw) and Y_4. The counting solution is

$$P(R_1 R_2 B_3 Y_4) = \frac{5 \cdot 4 \cdot 2 \cdot 3}{10 \cdot 9 \cdot 8 \cdot 7} = \tfrac{1}{42}.$$

The same calculation can be interpreted as an instance of (1.27):

$$\begin{aligned} P(R_1 R_2 B_3 Y_4) &= P(R_1)\, P(R_2 \,|\, R_1)\, P(B_3 \,|\, R_1 R_2)\, P(Y_4 \,|\, R_1 R_2 B_3) \\ &= \tfrac{5}{10} \cdot \tfrac{4}{9} \cdot \tfrac{2}{8} \cdot \tfrac{3}{7} = \tfrac{1}{42}. \end{aligned}$$

The interpretation of the numbers is for example that $P(B_3 \,|\, R_1 R_2) = \frac{2}{8}$ because after drawing two reds there are eight balls altogether two of which are black. △

Sometimes the information available is naturally conditional information. This example illustrates the use of the law of total probability.

Example 1.35. Suppose 90% of coins in circulation are fair and 10% are biased coins that give tails with probability 3/5. What is the probability that a flip of a randomly chosen coin yields tails?

Let F denote the event that the coin is fair, B that it is biased, and A the event that the flip yields tails. The information available gives us the probabilities of the two types of coins: $P(F) = \frac{9}{10}$ and $P(B) = \frac{1}{10}$, and the conditional probabilities of tails: $P(A \,|\, F) = \frac{1}{2}$ and $P(A \,|\, B) = \frac{3}{5}$. Then

$$P(A) = P(A \,|\, F)P(F) + P(A \,|\, B)P(B) = \tfrac{1}{2} \cdot \tfrac{9}{10} + \tfrac{3}{5} \cdot \tfrac{1}{10} = \tfrac{51}{100}.$$

△

Bayes' formula answers the following question: given that we observe an event A, what are the probabilities of mutually exclusive events B_k? In applications the events B_k might be competing explanations or hypotheses, and A is used as evidence to assess their probabilities.

Theorem 1.36. (Bayes' formula.) *Let $\{B_k\}$ be a countable partition of the sample space Ω. Then for any event A with $P(A) > 0$ and each k such that $P(B_k) > 0$,*

$$P(B_k \,|\, A) = \frac{P(AB_k)}{P(A)} = \frac{P(A \,|\, B_k)\, P(B_k)}{\sum_{i:\, P(B_i)>0} P(A \,|\, B_i)\, P(B_i)}. \tag{1.31}$$

Proof. The first equality in (1.31) is the definition of conditional probability. The second is the application of (1.29) to the denominator. □

When the partition in question has only two elements, namely D and D^c, Bayes' formula simplifies to the form

$$P(D \,|\, A) = \frac{P(A \,|\, D)P(D)}{P(A \,|\, D)P(D) + P(A \,|\, D^c)P(D^c)}. \tag{1.32}$$

Example 1.37. Suppose we have a medical test that detects a particular disease 96% of the time, but gives false positives 2% of the time. Assume that 0.5% of the population carries the disease. If a random person tests positive for the disease, what is the probability that they actually carry the disease?

Define the events $D = \{\text{person has the disease}\}$ and $A = \{\text{person tests positive}\}$. The problem statement gives the probabilities

$$P(A \,|\, D) = \tfrac{96}{100}\,, \quad P(A \,|\, D^c) = \tfrac{2}{100}\,, \quad \text{and} \quad P(D) = \tfrac{5}{1000}.$$

From Bayes' formula

$$P(D \,|\, A) = \frac{\frac{96}{100} \cdot \frac{5}{1000}}{\frac{96}{100} \cdot \frac{5}{1000} + \frac{2}{100} \cdot \frac{995}{1000}} = \frac{96}{494} \approx 0.194.$$

There are two competing explanations for the positive test: either the person has the disease, or the result is a false positive. The disease is rare compared to false positives, and consequently it is not the likeliest explanation of the positive test. By altering the inputs we see a different result. Suppose that, on the basis of an examination, a medical expert concludes that this person has a 50% chance to have the disease. Now we take $P(D) = 1/2$. With this updated $P(D)$ the calculation becomes

$$P(D \mid A) = \frac{\frac{96}{100} \cdot \frac{1}{2}}{\frac{96}{100} \cdot \frac{1}{2} + \frac{2}{100} \cdot \frac{1}{2}} = \frac{96}{98} \approx 0.980.$$

Now the disease is the very likely explanation of a positive test result. △

Example 1.38. We give yet another proof for the fact that if we roll a die repeatedly, then the probability of never seeing a 6 is equal to 0.

We consider the probability space of infinitely many die rolls introduced in Example 1.12. Let A denote the event that there are no sixes in any of the die rolls:

$$A = \{(s_1, s_2, \dots) : s_i \in \{1, 2, \dots, 5\} \text{ for all } i \geq 1\}.$$

(It is an instructive exercise to check that this is indeed an event in our probability space.) We would like to show that $P(A) = 0$. Let B be the event that the first die roll is a six. Then the law of total probability tells us that

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

We have $P(B) = \frac{1}{6}$ and $P(B^c) = \frac{5}{6}$. We also know that $P(A|B) = 0$, since if we condition on having a six on the first die roll then we already have a six in our sequence of die rolls. (In other words: $AB = \emptyset$.) Thus

$$P(A) = P(A|B^c)\tfrac{5}{6}.$$

What can we say about $P(A|B^c)$? If the first die roll is not a six, then we can just 'disregard' that die roll, and 'pretend' that we start the die rolls at the second roll. Hence the conditional probability of not seeing a six given that the first roll is not a six should be the same as the unconditional probability of never seeing a six: $P(A|B^c) = P(A)$. This leads to the equation

$$P(A) = P(A)\tfrac{5}{6}$$

and $P(A) = 0$.
Exercise 1.26 gives an outline for a rigorous proof of the last step. △

1.4. Independent events

Definition 1.39. Two events A and B are **independent** if

$$P(AB) = P(A)P(B). \tag{1.33}$$

△

The significance of the definition is best understood in light of conditional probability: if $P(B) > 0$, then A and B are independent if and only if $P(A|B) = P(A)$. In other words, knowledge about B happening has no influence on the probability of A.

Example 1.40. We flip a fair coin twice. Let A be the event that the first coin flip is heads and B the event that the second coin flip is tails. Our intuition tells us that A and B are independent, let us check this using the definition.

We use the probability space described in Example 1.13. The sample space is $\Omega = \{0,1\}^2$, $A = \{(0,1),(0,0)\}$, $B = \{(0,1),(1,1)\}$ and $AB = \{(0,1)\}$. Since we have equally likely outcomes, we have $P(A) = \frac{2}{4} = \frac{1}{2}$, $P(B) = \frac{2}{4} = \frac{1}{2}$ and $P(AB) = \frac{1}{4}$. This shows that $P(AB) = P(A)P(B)$, and hence A and B are indeed independent. △

Although sometimes independence might follow intuitively, the best way to verify it is to check the definition.

Example 1.41. Suppose that we flip a fair coin three times. Let A be the event that we have exactly one tails among the first two coin flips, B the event that we have exactly one tails among the last two coin flips and D the event that we get exactly one tails among the three coin flips. Show that A and B are independent, A and D are not independent, and B and D are also not independent.

We have $\#\Omega = 8$. The events A and B have four elements and D has three:

$$\begin{aligned} A &= \{(0,1,0),(0,1,1),(1,0,0),(1,0,1)\}, \\ B &= \{(0,0,1),(0,1,0),(1,0,1),(1,1,0)\}, \\ D &= \{(0,0,1),(0,1,0),(1,0,0)\}. \end{aligned}$$

The intersections are $AB = \{(0,1,0),(1,0,1)\}$, $AD = \{(0,1,0),(1,0,0)\}$, and $BD = \{(0,0,1),(0,1,0)\}$ which gives

$$\begin{aligned} P(AB) &= \tfrac{2}{8} = \tfrac{4}{8}\cdot\tfrac{4}{8} = P(A)P(B), \\ P(AD) &= \tfrac{2}{8} \neq \tfrac{4}{8}\cdot\tfrac{3}{8} = P(A)P(D), \\ P(BD) &= \tfrac{2}{8} \neq \tfrac{4}{8}\cdot\tfrac{3}{8} = P(B)P(D). \end{aligned}$$

△

Theorem 1.42. *Suppose that A and B are independent events. Then the same is true for each of these pairs: A^c and B, A and B^c, and A^c and B^c.*

Proof. If A and B are independent then $P(A)P(B) = P(AB)$. To prove the independence of A^c and B start from the identity $P(B) = P(A^cB) + P(AB)$ which follows from the fact that A^cB and AB are disjoint and their union is B. Then

$$\begin{aligned} P(A^cB) &= P(B) - P(AB) = P(B) - P(A)P(B) = (1-P(A))P(B) \\ &= P(A^c)P(B) \end{aligned}$$

which says that A^c and B are independent. The proof for the remaining two pairs follows the same way. □

Theorem 1.42 ties in with the notion that independence has to do with information. Knowing whether A happened is exactly the same as knowing whether A^c happened. (Simply because A happens if and only if A^c does not happen, and vice versa.) Thus if knowledge about A does not alter the probability of B, neither should knowledge about A^c.

Example 1.43. The events $\varnothing$ and Ω do not give any information about anything. Appropriately, they are independent of every other event A:

$$P(A\cap\varnothing) = P(\varnothing) = 0 = P(A)P(\varnothing) \quad \text{and} \quad P(A\cap\Omega) = P(A) = P(A)P(\Omega).$$

△

The definition of independence of more than two events requires that the product property hold for any subcollection of events.

Definition 1.44. Events $A_1, \dots, A_n$ are **independent** (or **mutually independent**) if for every collection $A_{i_1}, \dots, A_{i_k}$, where $2 \le k \le n$ and $1 \le i_1 < i_2 < \cdots < i_k \le n$,

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_k}). \tag{1.34}$$

△

Example 1.45. (Independence of three events.) To illustrate the definition, the requirement for the independence of three events A, B, and C is that these four equations all hold:

$$P(AB) = P(A)P(B), \quad P(AC) = P(A)P(C), \quad P(BC) = P(B)P(C),$$
$$\text{and} \quad P(ABC) = P(A)P(B)P(C).$$

△

Theorem 1.42 extends to more than two events.

Theorem 1.46. *Suppose events $A_1, \dots, A_n$ are mutually independent. Then for every collection $A_{i_1}, \dots, A_{i_k}$, where $2 \le k \le n$ and $1 \le i_1 < i_2 < \cdots < i_k \le n$, we have*

$$P(A^*_{i_1} A^*_{i_2} \cdots A^*_{i_k}) = P(A^*_{i_1}) P(A^*_{i_2}) \cdots P(A^*_{i_k}) \tag{1.35}$$

where each A^*_i can represent either A_i or A^c_i.

The proof of the theorem is left as Exercise 1.27. Exercise 1.28 shows that the converse of this theorem holds as well.

Example 1.47. Let us illustrate again with three independent events A, B, and C. Among the identities that Theorem 1.46 implies are for example these:

$$P(AB^c) = P(A)P(B^c), \quad P(A^cC^c) = P(A^c)P(C^c),$$
$$P(AB^cC) = P(A)P(B^c)P(C) \quad \text{and} \quad P(A^cB^cC^c) = P(A^c)P(B^c)P(C^c).$$

△

The following useful fact is a consequence of Theorem 1.46.

Fact 1.48. Suppose events $A_1, \dots, A_n$ are mutually independent. Let B be an event that can be expressed from $A_1, \dots, A_n$ using the usual set operations. Then $P(B)$ can be expressed as a function of the numbers $1, P(A_1), P(A_2), \dots, P(A_n)$ using addition, subtraction and multiplication.

Proof. We first prove that if the event B can be expressed from $A_1, \dots, A_n$ using the usual set operations, then it can also be written as the disjoint union of sets of the form $A^*_1 A^*_2 \cdots A^*_n$ where A^*_i is either A_i or A^c_i. If this is true, then the statement of the fact follows: the probability of such an event is equal to $\prod_{i=1}^n P(A^*_i)$ by Theorem 1.46, and $P(A^*_i) = P(A_i)$ or $1 - P(A_i)$. Since $P(B)$ is just the sum of these probabilities, we get that $P(B)$ can be expressed as a function of the numbers $1, P(A_1), P(A_2), \dots, P(A_n)$ using addition, subtraction and multiplication.

The first statement can be proven by induction on the number of set operations involved. Note first that there are 2^n different events of the form $A_1^* A_2^* \cdots A_n^*$, since each A_i* can be chosen to be one of two possible events. These events are mutually disjoint: if for a given j one has A_j and the other A_j^c in the jth position then the first event is a subset of A_j, the second is a subset of A_j^c, hence they are disjoint. Let us call an event of the form $A_1^* A_2^* \cdots A_n^*$ an *atomic event*.

To prove the base case of our induction we show that for any $1 \le j \le n$ the even A_j can be written as the union of atomic events. (This will be a disjoint union by the previous observation.) Consider all atomic events where $A_j^* = A_j$, we claim that the union of these is equal to A_j. Any such event is a subset of A_j (because it is the intersection of A_j with some other events), so this is true for their union as well. For any $\omega \in A_j$ we can set up an atomic event with $A_j^* = A_j$ that contains ω: for each $k \neq j$ we choose A_k^* as A_k if $\omega \in A_k$ and A_k^* if $\omega \notin A_k$. This shows that other direction. Hence A_j is the union of atomic events.

For the induction step we have to show that if we have events that are unions of atomic events then their complement, union, intersection will also be unions of atomic events. This can be checked using the definitions of the respective operations using the additional observation that Ω itself is the union of all 2^n atomic events. □

Events $A_1, \dots, A_n$ are said to be *pairwise independent* if for any two indices $i \neq j$ the events A_i and A_j are independent. This is a weaker constraint than mutual independence. The next example illustrates both mutually independent events and events that are pairwise independent but not mutually independent.

Example 1.49. Consider Example 1.41 of the three fair coin flips and define events A and B as in the example. Let C be the event that we have exactly one tails among the first and third coin flip. Example 1.41 showed that A and B are independent. The same argument shows that B and C are independent and also that A and C are independent. Thus A, B and C are pairwise independent. However these events are not mutually independent. We have seen that $AB = \{(0,1,0),(1,0,1)\}$. Consequently $ABC = \varnothing$ and $P(ABC) = 0$. But $P(A) = P(B) = P(C) = \frac{1}{2}$. Thus $P(ABC) \neq P(A)P(B)P(C)$.

To see an example of three independent events, for $i = 1,2,3$ let G_i be the event that the ith flip is tails. So for example

$$G_1 = \{(1,0,0),(1,0,1),(1,1,0),(1,1,1)\}.$$

Each $P(G_i) = \frac{1}{2}$. Each pairwise intersection has two outcomes: for example, $G_1G_2 = \{(1,1,0),(1,1,1)\}$. So $P(G_iG_j) = \frac{1}{4} = P(G_i)P(G_j)$ for $i \neq j$. Finally, for the triple intersection

$$P(G_1G_2G_3) = P\{(1,1,1)\} = \tfrac{1}{8} = (\tfrac{1}{2})^3 = P(G_1)P(G_2)P(G_3).$$

We have verified that G_1, G_2, G_3 are independent. △

The definition of mutual independence extends to infinite sequences of events, but it requires nothing new, just that any finite subcollection of events satisfies Definition 1.44.

Definition 1.50. Let $\{A_k\}_{k\in\mathbb{Z}_{>0}}$ be an infinite sequence of events in a probability space $(\Omega,\mathcal{F},P)$. Then events $\{A_k\}_{k\in\mathbb{Z}_{>0}}$ are independent if for each $n\in\mathbb{Z}_{>0}$, events $A_1,\dots,A_n$ are independent. △

Example 1.51. Consider the probability space of infinitely many fair coin flips described in Example 1.13. Let A_n the event that the nth coin flip is heads. We would guess that $\{A_k\}_{k\in\mathbb{Z}_{>0}}$ are independent, but let us check the definition.

We need to show that for each n the events $A_1,\dots,A_n$ are independent. For this we have to show that for any $1\le i_1<\cdots<i_k\le n$ we have

$$P(A_{i_1}\cdots A_{i_k})=\prod_{i=1}^k P(A_{i_k}).$$

We can compute this probability by counting favorable outcomes in the probability space of n coin flips with $\Omega=\{0,1\}^n$. Clearly, $\#A_{i_j}=2^{n-1}$, since we can choose the outcomes of the individual coin flips at $n-1$ positions. Thus $P(A_{i_j})=\frac{2^{n-1}}{2^n}=\frac{1}{2^n}$. The event $A_{i_1}\cdots A_{i_k}$ collects those outcomes where we have heads at positions $i_1,\dots,i_k$. We can choose the outcomes of the coin flips at the other $n-k$ positions freely, hence $\#(A_{i_1}\cdots A_{i_k})=2^{n-k}$ and $P(A_{i_1}\cdots A_{i_k})=\frac{2^{n-k}}{2^n}=2^{-k}$. But this shows that

$$2^{-k}=P(A_{i_1}\cdots A_{i_k})=\prod_{i=1}^k P(A_{i_k})=(\tfrac12)^k=2^{-k}.$$

This proves the independence of $A_1,\dots,A_n$ and also that of $\{A_k\}_{k\in\mathbb{Z}_{>0}}$. △

Sometimes independence is 'built into' the probability space (as in the previous example). Other times it is assumed from the setup of the problem.

Example 1.52. We have an unfair coin: the probability of getting tails when we flip it is $p\in(0,1)$. We flip the coin twice. Set up the probability space, if we assume that the events $A=\{\text{first flip is tails}\}$, $B=\{\text{second flip is tails}\}$ are independent.

We use the same sample space as in the case of the fair coin flips: $\Omega=\{0,1\}^2$. The σ-algebra of events is 2^Ω, so we only need to identify the probability measure P.

The events A and B are given by

$$A=\{(1,0),(1,1)\},\qquad B=\{(0,1),(1,1)\}.$$

By our assumptions $P(A)=P(B)=p$ and

$$P(A)P(B)=P(AB)=P\{(1,1)\}=p^2.$$

From this we get

$$\begin{aligned}P\{(1,0)\}&=P(A)-P(AB)=p-p^2=p(1-p)\\ P\{(0,1)\}&=P(B)-P(AB)=p-p^2=p(1-p)\\ P\{(0,0)\}&=1-P\{(1,0)\}-P\{(1,1)\}-P\{(0,1)\}\\ &=1-2p(1-p)-p^2=(1-p)^2.\end{aligned}$$

Another way to identify P is to use Theorem 1.46, and noting that

$$\{(1,0)\}=AB^c,\qquad \{(0,1)\}=A^cB,\qquad \{(0,0)\}=A^cB^c.$$

△

The following is an intuitively obvious fact, but we cannot prove it in general with the techniques available to us. Small special cases can be handled, as in the example below.

Theorem 1.53. *Let $\{A_k\}_{k\geq 1}$ be a finite or infinite sequence of independent events. Let $0 = k_0 < k_1 < \cdots < k_n$ be integers. Let $B_1, \dots, B_n$ be events constructed from the A_ks so that, for each $j = 1, \dots, n$, B_j is formed by applying set operations to $A_{k_{j-1}+1}, \dots, A_{k_j}$. Then the events $B_1, \dots, B_n$ are independent.*

The point of the formulation is that no two different B_js use the same A_k. The next example and Exercise 1.35 illustrate.

Example 1.54. Let A, B and C be independent events. Show that the two events A and $B^c \cup C$ are independent. To show this, we use inclusion-exclusion and the fact that when events are independent, the product rule of independence continues to hold if some events are replaced with complements.

$$\begin{aligned}
P\big(A \cap (B^c \cup C)\big) &= P\big((A \cap B^c) \cup (A \cap C)\big) \\
&= P(A \cap B^c) + P(A \cap C) - P(A \cap B^c \cap C) \\
&= P(A)P(B^c) + P(A)P(C) - P(A)P(B^c)P(C) \\
&= P(A)\big[P(B^c) + P(C) - P(B^c)P(C)\big] \\
&= P(A)P(B^c \cup C).
\end{aligned}$$

You can imagine how tedious the proof would be for an example with more events. However, this is not inherently so, but only because we do not have the right mathematical tools at our disposal. △

Independence of random variables can be defined via the independence of events related to those random variables.

Definition 1.55. Let $X_1, X_2, \dots, X_n$ be random variables on the same probability space. The random variables $X_1, \dots, X_n$ are mutually independent if for all choices of $c_1, c_2, \dots, c_n \in \mathbb{R}$ the events $A_k = \{\omega : X_k(\omega) \leq c_k\}$ are mutually independent.

Let $X_1, X_2, \dots$ be a sequence of random variables on the same probability space. The random variables $X_1, X_2, \dots$ are mutually independent if for any $n \geq 1$ the random variables $X_1, \dots, X_n$ are mutually independent.

We will have a closer look at random variables (and independent random variables) in the next chapter. In the meantime, you could check that if $X_k(s_1, s_2, \dots) = s_k$ denotes the outcome of the kth die roll in Example 1.12 then the random variables $X_1, X_2, \dots$ are mutually independent.

Conditional independence.

When the definition of independence is applied to a conditioned probability measure $P(\cdot \,|\, B)$, we call it conditional independence.

Definition 1.56. Let $A_1, \dots, A_n$ and B be events on $(\Omega, \mathcal{F}, P)$ and assume $P(B) > 0$. Then events $A_1, \dots, A_n$ are **conditionally independent, given B,** if for every collection $A_{i_1}, \dots, A_{i_k}$, where $2 \le k \le n$ and $1 \le i_1 < i_2 < \cdots < i_k \le n$,

$$P(A_{i_1} A_{i_2} \cdots A_{i_k} \,|\, B) = \prod_{j=1}^{k} P(A_{i_j} \,|\, B). \tag{1.36}$$

△

Example 1.57. (Continuation of Example 1.35.) Suppose 90% of coins in circulation are fair and 10% are biased coins that give tails with probability 3/5. I have a random coin and I flip it twice. Denote by A_1 the event that the first flip yields tails and by A_2 the event that the second flip yields tails. Should the two flips be independent, in other words, should we expect to have $P(A_1A_2) = P(A_1)P(A_2)$?

Let F denote the event that the coin is fair and B that it is biased. The information above gives us the probabilities of the two types of coins: $P(F) = \frac{9}{10}$ and $P(B) = \frac{1}{10}$, and the conditional probabilities of tails:

$$P(A_1 \,|\, F) = P(A_2 \,|\, F) = \tfrac{1}{2}, \qquad P(A_1 \,|\, B) = P(A_2 \,|\, B) = \tfrac{3}{5}.$$

Above we made the natural assumption that *for a given coin*, the probability of tails does not change between the first and second flip. Then we compute, for both $i = 1$ and 2,

$$P(A_i) = P(A_i \,|\, F)P(F) + P(A_i \,|\, B)P(B) = \tfrac{1}{2} \cdot \tfrac{9}{10} + \tfrac{3}{5} \cdot \tfrac{1}{10} = \tfrac{51}{100}.$$

To compute the probability of two tails, we need to make an assumption: the successive flips *of a given coin* are independent. This gives us the conditional independence:

$$P(A_1A_2 \,|\, F) = P(A_1 \,|\, F)\, P(A_2 \,|\, F) \quad \text{and} \quad P(A_1A_2 \,|\, B) = P(A_1 \,|\, B)\, P(A_2 \,|\, B).$$

Then we apply again the law of total probability:

$$\begin{aligned} P(A_1A_2) &= P(A_1A_2 \,|\, F)\, P(F) + P(A_1A_2 \,|\, B)\, P(B) \\ &= P(A_1 \,|\, F)\, P(A_2 \,|\, F)\, P(F) + P(A_1 \,|\, B)\, P(A_2 \,|\, B)\, P(B) \\ &= \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{9}{10} + \tfrac{3}{5} \cdot \tfrac{3}{5} \cdot \tfrac{1}{10} = \tfrac{261}{1000}. \end{aligned}$$

Now $\frac{261}{1000} \ne \left(\frac{51}{100}\right)^2$ and we conclude that $P(A_1A_2) \ne P(A_1)P(A_2)$. In other words, A_1 and A_2 are *not* independent without the conditioning. The intuitive reason is that the first flip gives us information about the coin we hold, and thereby alters our expectations about the second flip. Exercise 1.33 develops these ideas further. △

1.5. Further mathematical issues ♣

Probability measures and σ-algebras.

In discrete sample spaces Ω we can typically define a probability $P(A)$ for all subsets A of Ω. Consequently in the definition of the probability space $(\Omega, \mathcal{F}, P)$, the σ-algebra $\mathcal{F}$ can be the power set of Ω. This convenience breaks down as we move from countable to uncountable sample spaces. In particular, it is impossible to define a probability measure P on *all* subsets of the unit interval $[0, 1]$ so that the countable additivity axiom (1.2) holds and $P(I) =$ length of I for all intervals

$I \subset [0,1]$. (A 3/4 page proof of this which uses nothing beyond an equivalence relation appears in the beginning of Chapter 1 in [Fol99].)

A consequence is that in general we cannot expect to construct probability spaces that satisfy Definition 1.2 and have $P(A)$ defined for all subsets A of Ω. It becomes necessary to find σ-algebras that contain enough sets to be useful for calculations but not so many that a probability measure P cannot be defined. The right notion for theory and practice on metric spaces $\mathcal{X}$ is the *Borel σ-algebra* $\mathcal{B}_{\mathcal{X}}$ which is by definition the σ-algebra generated by the open subsets of $\mathcal{X}$. Technically $\mathcal{B}_{\mathcal{X}}$ is defined as the intersection of all the σ-algebras that contain the open subsets of $\mathcal{X}$. Then $\mathcal{B}_{\mathcal{X}}$ is the smallest σ-algebra on $\mathcal{X}$ that contains all the open subsets.

The reader may find the analogy with linear algebra helpful. A set of vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$ in a vector space generates or *spans* the vector subspace $V = \operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$. V can be equivalently defined as the intersection of all the vector subspaces that contain the vectors $\mathbf{x}_1, \dots, \mathbf{x}_m$. V is also then the smallest vector subspace that contains $\mathbf{x}_1, \dots, \mathbf{x}_m$. However, this analogy does not go very far. Every vector in V can be written as a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_m$. But no simple representations exist for all Borel subsets of uncountable metric spaces.

On the Borel σ-algebra of $[0,1]$ there exists the *Lebesgue measure* P that satisfies $P(I) = \text{length of } I$ for all intervals $I \subset [0,1]$. This construction is in every textbook on measure theory. More generally, if Ω is a bounded Borel subset of $\mathbb{R}^d$, there exists a probability space $(\Omega, \mathcal{F}, P)$ such that $\mathcal{F} = \mathcal{B}_\Omega$ is the Borel σ-algebra of Ω and for each $A \in \mathcal{F}$, $P(A) = \operatorname{vol}(A)/\operatorname{vol}(\Omega)$ where the "volume" $\operatorname{vol}(A)$ has to be interpreted in the way appropriate for the dimension: in $d = 1$ it is length, in $d = 2$ area, in $d = 3$ actual physical volume, and in $d > 3$ a generalized volume. This is the probability space for picking a uniformly random point from the set Ω.

The key point for practice is that the Borel σ-algebra contains all the sets that one normally wants to work with. Constructing non-Borel sets requires some ingenuity. This is the reason that we can study probability and safely set the measure-theoretic details aside.

Exercises

Exercise 1.1. Let $\Omega = \mathbb{R}$. Fix a point $z \in \mathbb{R}$ and define for each subset $A \subset \mathbb{R}$,

$$P(A) = \begin{cases} 1, & \text{if } z \in A \\ 0, & \text{if } z \notin A. \end{cases}$$

Let $\mathcal{F}$ be the power set of $\mathbb{R}$, that is, the collection of all subsets of $\mathbb{R}$. Verify that $(\Omega, \mathcal{F}, P)$ satisfies the axioms of a probability space.

A probability measure such as P in this exercise that is concentrated on a single point is called a *degenerate* probability measure. Since the outcome is z with probability 1, there is really no randomness in the outcome at all.

Exercises 1.2–1.6 practice three basic sampling situations where outcomes are equally likely and hence probabilities can be computed by counting.

Exercise 1.2. An urn contains three balls: one green, one red and one white ball. A random ball is drawn repeatedly, its color observed, and the ball put back into the urn before the next draw. This is called *sampling with replacement.* In parts (a)–(c) below a fixed finite number of draws are made. In parts (d)–(g) infinitely many draws are made.

(a) Let $n \in \mathbb{Z}_{>0}$. A ball is drawn n times and the n colors recorded in order. For example, if $n = 3$, a possible outcome is $(\texttt{r}, \texttt{g}, \texttt{r}) = (\texttt{red}, \texttt{green}, \texttt{red})$. Assume that each ordered n-tuple of colors is equally likely. Give a complete description of the probability space $(\Omega, \mathcal{F}, P)$ that models this experiment.

(b) Suppose n draws are made. Define the events

$$A_1 = \{\text{the green ball is seen at least once}\}$$

and

$$A_2 = \{\text{the green ball is seen at least twice}\}.$$

Calculate the probabilities $P(A_1)$ and $P(A_2)$.

(c) Suppose n draws are made. Let C_n be the event that not all colors were seen. Find $P(C_n)$.

(d) Let C be the event that in infinitely many draws there is some color that is never seen. Find $P(C)$.

(e) A ball is drawn repeatedly until the green ball or the red ball appears. Let D be the event that the red ball appears before the green ball. Compute the probability $P(D)$ by decomposing the event D according to the number of draws needed to see the first red ball.

(f) Let U be the event that the green ball appears on an even draw before the first time when the red ball appears on an odd draw. Find $P(U)$.

(g) Fix two positive integers $m < n$. Let $B_{m,n}$ be the event that in n draws, the green ball does not appear after the mth draw. Let $B_{m,\infty}$ be the event that in infinitely many draws, the green ball does not appear after the mth draw. Let E be the event that in infinitely many draws, the green ball appears infinitely many times. Find the probabilities $P(B_{m,n})$, $P(B_{m,\infty})$, and $P(E)$.

Exercise 1.3. An urn contains five balls labeled 1 through 5. Three balls are removed from the urn one by one randomly, and their numbers recorded in order. No balls are put back in the urn. Assume that all outcomes of the experiment are equally likely. This is called *sampling without replacement.* The difference with Exercise 1.2 is that now no ball can be drawn more than once.

(a) Give a description of the probability space $(\Omega, \mathcal{F}, P)$ appropriate for this experiment.

(b) Let A be the event that the second ball drawn is ball 4 or 5. Find $P(A)$.

(c) Let B be the event that the sample of three balls contains ball 1 or ball 2 or both. Find $P(B)$.

(d) Let C be the event that the three recorded numbers come in increasing order. Find $P(C)$.

Exercise 1.4. A committee of three people is chosen uniformly at random from a group of seven people. Mathematically, a committee is a *set* of people that does not come with any order. This is another instance of *sampling without replacement*.

(a) Give a description of the probability space $(\Omega, \mathcal{F}, P)$ appropriate for this experiment.

(b) Among the seven people are Judy and Phil. Let A be the event that Judy is included in the committee but Phil is left out. Find $P(A)$.

(c) Suppose now that the group of seven consists of four women and three men, and the committee must be composed of two women and one man. Give a description of the probability space. Let A be the same event as above. Compute the probability $P(A)$.

Exercise 1.5. The Powerball lottery has the following rules. There are 69 white balls (numbered from 1 to 69) and 26 red balls (numbered from 1 to 26). At each drawing five white balls and a red ball are chosen uniformly at random. The order of the white balls is not important: the white numbers are listed in decreasing order after the drawing, together with the red number.

(a) Give a description of the probability space $(\Omega, \mathcal{F}, P)$ appropriate for a single Powerball drawing.

(b) Find the probability that the picked six numbers are all different.

(c) Find the probability that the number 13 is picked (either as white or red or both).

Exercise 1.6. In the lottery of Example 1.7, let B be the event that 13 and 14 are included but 15 is not. Calculate the probability $P(B)$ with (a) ordered samples and (b) unordered samples.

Exercise 1.7. Pick a uniformly chosen random point inside a square of side length 4 and draw a circle of radius 1 around the point. Find the probability that the circle lies entirely inside the square.

Exercise 1.8. Let A and B be two events on $(\Omega, \mathcal{F}, P)$. Suppose $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{3}$.

(a) Show that $\frac{1}{12} \le P(A \cap B) \le \frac{1}{3}$. Give examples to show that both extreme values $P(A \cap B) = \frac{1}{12}$ and $P(A \cap B) = \frac{1}{3}$ can be achieved.

(b) Do the same for $P(A \cup B)$. Namely, find upper and lower bounds for $P(A \cup B)$ that are valid for all events A and B that satisfy $P(A) = \frac{3}{4}$ and $P(B) = \frac{1}{3}$, and give examples where the extreme values are achieved.

Exercise 1.9. We have two red, two white and two green marbles in an urn. We pick them one by one out of the urn and record their colors. (Note: this is sampling without replacement.) Find the probability that at some point we will pick the same color back to back. (E.g. this happens when we get the sequence red, white, white, green, red, green, but also if we get red, red, white, white, green, green.)

Exercise 1.10. Give an example of an uncountable collection of events where each event has probability one but their intersection has probability zero.

Hint. Think of Example 1.9.

Exercise 1.11. Is it possible to choose uniformly at random a rational number from the unit interval $[0,1]$?

Exercise 1.12. Consider the following experiment. We have an urn with two marbles numbered 1 and 2. We pick a marble randomly, write down its number and return it to the urn. Then we add a marble with the number 3 to the urn, choose one of the three marbles randomly, record its number and return it to the urn. We repeat this over and over: so before the $(k+1)$st pick we add a marble with with the number $k+2$ in the urn (so that it contains the numbers $1,\dots,k+2$), choose a marble randomly, record its number and return it to the urn. At the end of the experiment we have an infinite sequence of integers.

(a) Set up the probability space using the strategy described in Examples 1.12 and 1.13. You may assume that for any given k the outcomes for the first k picks are equally likely.

(b) Show that with probability one the marble with the number 1 will be picked at some point.
Hint. First find the probability that for a given k the marble with the number 1 is picked in the kth pick at the first time.

Exercise 1.13. Let $\{A_k\}_{k\in\mathbb{Z}_{\geq 0}}$ be a sequence of events such that $A_k \nearrow A$ and $A_0 = \varnothing$. Show that

(a) the events $\{A_k \setminus A_{k-1}\}_{k\in\mathbb{Z}_{>0}}$ are pairwise disjoint, and

(b) $\displaystyle\bigcup_{k=1}^{\infty} A_k = \bigcup_{k=1}^{\infty}(A_k \setminus A_{k-1})$.

Exercise 1.14. Show that $A_k \searrow A$ implies $P(A_k) \to P(A)$ from the axioms. Do not appeal to Theorem 1.25, and do not simply apply de Morgan's law and then tap into the proof of the case $A_k \nearrow A$ in Theorem 1.25.

Hint. Sketch a picture like Figure 1.

Exercise 1.15. We have two urns. The first urn contains two balls labeled 1 and 2. The second urn contains three balls labeled 3, 4 and 5. We choose one of the urns at random (with equal probability) and then sample one ball (uniformly at random) from the chosen urn. What is the probability that we picked the ball labeled 5?

Exercise 1.16. When Alice spends the day with the babysitter, there is a 0.6 probability that she turns on the TV and watches a show. Her little sister Betty cannot turn the TV on by herself. But once the TV is on, Betty watches with probability 0.8. Tomorrow the girls spend the day with the babysitter.

(a) What is the probability that both Alice and Betty watch TV tomorrow?

(b) What is the probability that Betty watches TV tomorrow?

(c) What is the probability that only Alice watches TV tomorrow?

Hint. Define events precisely and use the product rule and the law of total probability.

Exercise 1.17. Let $P(A \cap B) > 0$. Show that conditioning on A, followed by conditioning on B, is the same as conditioning on $A \cap B$. Explicitly: if probability measure Q is defined by $Q(C) = P(C \,|\, A)$, then $Q(C \,|\, B) = P(C \,|\, A \cap B)$.

Exercise 1.18. Let B be an event on the probability space $(\Omega, \mathcal{F}, P)$ such that $P(B) > 0$. Show that, as a function of the event $A \in \mathcal{F}$, $P(A|B)$ satisfies the axioms of a probability measure.

Exercise 1.19. Fix a parameter $0 \le \theta \le 1$. Bob rolls a die repeatedly in the hopes of seeing a six. However, after each failure to see a six he gives up with probability $1 - \theta$ and decides to try again with probability θ. What is the probability that Bob never sees a six?

Exercise 1.20. An insurance company has two types of customers, careful and reckless. A careful customer has an accident during the year with probability 0.01. A reckless customer has an accident during the year with probability 0.04. 80% of the customers are careful and 20% of the customers are reckless. Suppose a randomly chosen customer has an accident this year. What is the probability that this customer is one of the careful customers?

Exercise 1.21. I have three coins in my pocket, labeled a, b and c. The probabilities with which the coins give tails are $p_a, p_b, p_c \in (0, 1)$, respectively. I grab one coin out of the three at random. I flip it twice and count the number of tails.

(1) Define a probability space that models the choice of coin and the number of tails.

(2) Given that both flips were tails, what is the probability that I picked coin a?

Exercise 1.22. Suppose a family has 2 children of different ages. We assume that all combinations of boys and girls are equally likely.

(a) Formulate precisely the sample space and probability measure that describes the genders of the two children in the order in which they are born.

(b) Suppose we learn that there is a girl in the family. (Precisely: we learn that there is at least one girl.) What is the probability that the other child is a boy?

(c) Suppose we see the parents with a girl, and the parents tell us that this is their youngest child. What is the probability that the older child we have not yet seen is a boy?

Exercise 1.23. Assume that $\frac{1}{3}$ of all twins are identical twins. You learn that Miranda is expecting twins, but you have no other information.

(a) Find the probability that Miranda will have two girls.

(b) You learn that Miranda gave birth to two girls. What is the probability that the girls are identical twins?

Explain any assumptions you make.

Exercise 1.24. A crime has been committed in a town of 100,000 inhabitants. The police are looking for a single perpetrator, believed to live in town. DNA evidence is found on the crime scene. Kevin's DNA matches the DNA recovered from the crime scene. According to DNA experts, the probability that a random person's

DNA matches the crime scene DNA is 1 in 10,000. Before the DNA evidence, Kevin was no more likely to be the guilty person than any other person in town. What is the probability that Kevin is guilty after the DNA evidence appeared?
Hint. Reason as in Example 1.37.

Exercise 1.25 (Prisoner's paradox)**.** Three prisoners A, B and C have been sentenced to die tomorrow. The king has chosen one of the three uniformly at random to be pardoned tomorrow, while the two unlucky ones head for the gallows. The guard already knows who is to be pardoned. Prisoner A begs the guard to name someone other than A himself who will be executed. He cajoles, "Even if you tell me which one of B and C will be executed, I will not have gained any knowledge because I know already that at least one of them will die tomorrow." The guard is persuaded and reveals that B is to die tomorrow.

(a) After receiving this information, does A still have probability $\frac{1}{3}$ of being pardoned?

(b) Prisoner A whispers his new information to prisoner C. Prisoner C learned conditional probability before turning to a life of crime and is now hopeful. What is his new probability of being pardoned?

Hint. Use events $A = \{A \text{ is pardoned}\}$, similarly for B and C, and $D = \{\text{the guard names } B \text{ as one to die}\}$. Compute $P(A|D)$ and $P(C|D)$. Your answer will depend on the quantity $p = P(D|A) =$ the probability that the guard names B if both B and C are to die. Interpret in particular the special cases $p = 0$, 1 and $\frac{1}{2}$.

Exercise 1.26. Provide a rigorous proof for the last step in Example 1.38. Consider the probability space of infinitely many die rolls. Let A be the event that there are no sixes, and B the event that the first roll is a six. Show that if $P(A|B^c) = P(A)$.
Hint. Here is a possible outline:
- Show that if C is an event that does not depend on the first die roll then $P(C|B^c) = P(C)$.
- Use the previous step for the event that there are no sixes starting with the second die roll.

Exercise 1.27. Prove Theorem 1.46.
Hint. Show first that it is enough to prove (1.35) for $k = n$, i.e. when $i_j = j$ for $1 \le j \le n$.

Exercise 1.28. Prove the following converse of Theorem 1.46. Suppose that $A_1, \dots, A_n$ are events for which

$$P(A_1^* A_2^* \cdots A_n^*) = P(A_1^*)P(A_2^*) \cdots P(A_n^*)$$

holds for all 2^n possible choices $A_1^*, \dots, A_n^*$. (Each A_i^* can be either A_i or A_i^c.) Then $A_1, \dots, A_n$ are mutually independent.
Hint. Use induction.

Exercise 1.29. Let A and B be two disjoint events. Under what conditions are A and B independent?

Exercise 1.30. Suppose that a person's birthday is a uniformly random choice from the 365 days of a year (leap years are ignored), and one person's birthday is independent of the birthdays of other people. Alex, Betty and Conlin are comparing birthdays. Define these three events:

$$\begin{aligned} A &= \{\text{Alex and Betty have the same birthday}\} \\ B &= \{\text{Betty and Conlin have the same birthday}\} \\ C &= \{\text{Conlin and Alex have the same birthday}\}. \end{aligned}$$

(a) Are events A, B and C pairwise independent? (See the definition of pairwise independence on page 32 and Example 1.49 for illustration.)

(b) Are events A, B and C independent?

Exercise 1.31. Two towns are connected by a daily bus and by a daily train that go through a valley. On any given day one or more of the following three mutually independent events may happen: (i) the bus breaks down with probability $2/10$, (ii) the train breaks down with probability $1/10$, and (iii) a storm closes down the valley and cuts off both the bus and the train with probability $1/20$. What is the probability that travel is possible between the two towns tomorrow?

Exercise 1.32. This problem investigates the properties of conditional independence. Let A, B, F be three events that all have positive probability. By definition, A and B are conditionally independent, given F, if

$$P(AB \mid F) = P(A \mid F)P(B \mid F). \tag{1.37}$$

(a) Is the independence of A, B and F a sufficient condition for (1.37)?

(b) Is the independence of A, B and F a necessary condition for (1.37)?

(c) Show that (1.37) is equivalent to $P(B \mid AF) = P(B \mid F)$.

Exercise 1.33. As in Example 1.57, assume that 90% of the coins in circulation are fair, and the remaining 10% are biased coins that give tails with probability $3/5$. I hold a randomly chosen coin and begin to flip it.

(a) After one flip that results in tails, what is the probability that the coin I hold is a biased coin? After two flips that both give tails? After n flips that all come out tails?

(b) After how many straight tails can we say that with 90% probability the coin I hold is biased?

(c) After n straight tails, what is the probability that the next flip is also tails?

(d) Suppose we have flipped a very large number of times (think number of flips n tending to infinity), and each time gotten tails. What are the chances that the next flip again yields tails?

Exercise 1.34. Suppose that 1% of the employees of a certain company use illegal drugs. This company performs random drug tests that return positive results 99% of the time if the person is a drug user. However, it also has a 2% false positive rate. The results of the drug test are known to be independent from test to test for a given person.

(a) Steve, an employee at the company, has a positive test. What is the probability he is a drug user?

(b) Knowing he failed his first test, what is the probability that Steve will fail his next drug test?

(c) Steve just failed his second drug test. Now what is the probability he is a drug user?

Exercise 1.35. Let $\{A_k\}$ be a sequence of independent events. Using *only* the definition of independence, the additivity of probability, the inclusion-exclusion identity, and set operations, show that the events $A_1 A_2^c$ and $A_3 \cup A_4$ are independent.

Exercise 1.36. Let $G_i = \{i\text{th flip is tails}\}$ in Example 1.49 of three fair coin flips. Construct an event A by applying set operations to G_1 and G_2, and an event B by applying set operations to G_2 and G_3, and so that A and B are not independent.

Exercise 1.37. Consider the probability space constructed in Example 1.13 for flipping a fair coin infinitely many times.

(a) Find an event A in the probability space with $P(A) = \frac{5}{8}$.

(b) Find an event B in the probability space with $P(B) = \frac{1}{3}$.

(c) Given $0 < p < 1$ construct an event C in the probability space with $P(C) = p$.

The solution of part (c) gives a way to simulate a probability p event for any p with a single coin!

Exercises for Section 1.5 ♣.

Exercise 1.38. Let Ω be at most countable. Define a metric d on Ω by

$$d(\omega, \omega') = \begin{cases} 1, & \text{if } \omega \neq \omega' \\ 0, & \text{if } \omega = \omega'. \end{cases}$$

Show that the Borel σ-algebra of Ω is the power set of Ω.

Exercise 1.39. Let Ω be the sample space $\{0,1\}^{\mathbb{Z}_{>0}}$ defined in (1.10) for infinitely many coin tosses. Define a *product metric* on Ω by

$$d(\omega, \omega') = \sum_{k=1}^{\infty} 2^{-k}\, I_{\{s_k \neq s_k'\}}$$

where $\omega = (s_k)_{k \in \mathbb{Z}_{>0}}$ and $\omega' = (s_k')_{k \in \mathbb{Z}_{>0}}$ are two sequences in Ω and the indicator $I_{\{s_k \neq s_k'\}}$ equals 1 if the coordinates s_k and s_k' disagree, zero otherwise. Show that event $A_{n,\mathbf{t}}$ of (1.11) is an open subset of Ω. The point is that then we know that $A_{n,\mathbf{t}}$ is also a Borel subset of Ω.

Chapter 2

Random variables and probability distributions

Often we are interested in some numerical value associated to the outcome of a random experiment. For example, when rolling a pair of dice it may be the sum of the two dice that is of interest, as for example in the game of Monopoly. Or, in Example 1.11 of a randomly thrown dart, we may want to know the distance from the random point to the center of the dart board. The mathematical way to attach numerical values to outcomes of a random experiment is to define a function from Ω into the real line $\mathbb{R}$. These functions are called *random variables*. Random variables generalize naturally to random vectors. The randomness of the values of random variables and vectors are described by their probability distributions.

2.1. Random variables

Recall from Definition 1.3 that a random variable on a probability space $(\Omega, \mathcal{F}, P)$ is a real valued function $\Omega \to \mathbb{R}$.

The conventions concerning random variables differ from the ones used for functions in other areas of mathematics. There is the term itself: a random variable is a function and not a variable. Instead of lower case letters such as f, g and h often used to denote functions, random variables are typically denoted by capital letters such as X, Y and Z. The value of a random variable X at sample point ω is $X(\omega)$.

There is a simple way to encode an events as a random variable.

Definition 2.1. Let B be an event in a probability space. Then the *indicator function* (or indicator random variable) of B is defined as the function

$$I_B(\omega) = \begin{cases} 1, & \text{if } \omega \in B \\ 0, & \text{if } \omega \notin B. \end{cases} \tag{2.1}$$

Note that I_B is a random variable.

The value $I_B(\omega)$ of an indicator random variable records whether ω lies in B or not. Indicator random variables will provide a useful tool for various computations.

Example 2.2. Consider the roll of two dice. The dice are distinguished from each other. A sample point is a pair $\omega = (i, j)$ where i is the outcome of the first die and j is the outcome of the second die. Following Example 1.5, the sample space is

$$\Omega = \{\omega = (i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$$

with equal probability for each outcome: $P\{\omega\} = \frac{1}{36}$ for each $\omega \in \Omega$.

We introduce four random variables on this probability space: X_1 is the outcome of the first die, X_2 is the outcome of the second die, S is the sum of the two dice, and Z is the number of rolls that give one or two. Here are formulas for these random variables as functions of a sample point $(i, j) \in \Omega$.

$$X_1(i, j) = i, \quad X_2(i, j) = j, \quad \text{and} \quad S(i, j) = X_1(i, j) + X_2(i, j) = i + j.$$

To give Z an explicit formula, we use indicator variables:

$$Z(i, j) = I_{\{1,2\}}(i) + I_{\{1,2\}}(j).$$

To illustrate with a particular sample point, suppose the first die is a six and the second die is a two. Then the realization of the experiment is $(6, 2)$ and the random variables take on the values

$$\begin{aligned} &X_1(6, 2) = 6, \quad X_2(6, 2) = 2, \quad S(6, 2) = 6 + 2 = 8, \\ &\qquad \text{and} \quad Z(6, 2) = I_{\{1,2\}}(6) + I_{\{1,2\}}(2) = 0 + 1 = 1. \end{aligned}$$

In random variable notation, computing the probability that the sum of the two dice is 8 goes as follows:

$$\begin{aligned} P\{S = 8\} &= \sum_{(i,j):i+j=8} P\{X_1 = i,\, X_2 = j\} \\ &= P\{X_1 = 2,\, X_2 = 6\} + P\{X_1 = 3,\, X_2 = 5\} + P\{X_1 = 4,\, X_2 = 4\} \\ &\qquad + P\{X_1 = 5,\, X_2 = 3\} + P\{X_1 = 6,\, X_2 = 2\} \\ &= 5 \cdot \tfrac{1}{36} = \tfrac{5}{36}. \end{aligned} \tag{2.2}$$

△

As seen above, random variables provide a convenient language for expressing events. Any event stated in terms of a random variable X is of the form "X takes certain values". In mathematical terms the completely general form of such an event is $\{\omega \in \Omega : X(\omega) \in B\}$, which is the set of sample points ω such that $X(\omega) \in B$, and B is some subset of $\mathbb{R}$. The full-fledged set expression $\{\omega \in \Omega : X(\omega) \in B\}$ is abbreviated to $\{X \in B\}$ and expressed in English as "X lies in B", or the inverse image of B with respect to X. The event is sometimes denoted as $X^{-1}(B)$ (even if X itself does not have an inverse function). For example, in (2.2) above, $\{S = 8\}$ is the abbreviated version of the expression $\{(i, j) \in \Omega : S(i, j) = 8\}$, and we could have used the notation $S^{-1}(\{8\})$ as well.

Example 2.3. Let N be the number of rolls of a fair die needed to see the first six (already looked at in Example 1.8). To rigorously define N as a random variable, we use the sample space of infinitely many fair die rolls:

$$\Omega = \{\omega = (s_k)_{1\le k<\infty} : s_k \in \{1,2,3,4,5,6\}\ \forall k \in \mathbb{Z}_{>0}\} \tag{2.3}$$

as described in Example 1.12. Define

$$N(\omega) = \inf\{k \in \mathbb{Z}_{>0} : s_k = 6\} \quad \text{for} \quad \omega = (s_k)_{1\le k<\infty}. \tag{2.4}$$

In plain English, $N(\omega)$ is the first index k such that $s_k = 6$. To take an example, if $\omega = (1,3,1,2,6,5,1,\dots)$, then $N(\omega) = 5$. By convention, the infimum of the empty set is infinity: $\inf\varnothing = \infty$. If six never appears in the sequence ω, then the set $\{k \in \mathbb{Z}_{>0} : s_k = 6\}$ is empty, and the value of N is $N(\omega) = \infty$.

The random variables X_1, X_2, S and Z of Example 2.2 can also be defined on the sample space (2.3) of infinitely many die rolls: for any sequence $\omega = (s_k)_{1\le k<\infty}$,

$$\begin{aligned} X_1(\omega) = s_1, \quad X_2(\omega) = s_2, \quad S(\omega) &= s_1 + s_2 \\ \text{and} \quad Z(\omega) &= I_{\{1,2\}}(s_1) + I_{\{1,2\}}(s_2). \end{aligned} \tag{2.5}$$

For $\omega = (1,3,1,2,6,5,1,\dots)$ we have the values

$$X_1(\omega) = 1, \quad X_2(\omega) = 3, \quad S(\omega) = 4 \quad \text{and} \quad Z(\omega) = 1.$$

△

Example 2.4. As in Example 1.11, pick a uniform random point from the disk of radius r_0 centered at the origin. The sample space is $\Omega = \{\omega = (x,y) : x^2+y^2 \le r_0^2\}$.

For each sample point $\omega = (x,y) \in \Omega$, define

$$R(x,y) = \sqrt{x^2+y^2}.$$

The random variable R represents the distance from the origin to the random point. △

We turn to describe the randomness of the values of a random variable.

Probability distribution of a random variable.

Through the probabilities of events of type $\{X \in B\}$, a random variable induces a probability measure on the real line.

Definition 2.5. Let X be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The **probability distribution** of X is the probability measure μ on $\mathbb{R}$ defined by

$$\mu(B) = P(X \in B) \tag{2.6}$$

for Borel subsets B of $\mathbb{R}$.

Remark 2.6. The reason for the use of Borel subsets is the following. If we want to define a probability measure on $\mathbb{R}$ then it is natural to include all intervals in the set of events. But if we do that, then we have to include all sets that can be generated from the intervals using the operations described in the definition of the σ-algebra in Definition 1.2, and this means that we have to include all Borel sets. (See Example 1.9.)

The probability $P(X \in B)$ is not defined unless the event $\{X \in B\}$ is a member of $\mathcal{F}$. A function $X : \Omega \to \mathbb{R}$ for which $\{X \in B\} \in \mathcal{F}$ for all Borel subsets B of $\mathbb{R}$ is called a *measurable function* on Ω. According to Definition 1.3 a random variable is function from Ω into $\mathbb{R}$ for which $\{X \in (-\infty, c]\} \in \mathcal{F}$ for each c. It can be shown that this condition implies that $\{X \in B\} \in \mathcal{F}$ for all Borel subsets B of $\mathbb{R}$. Thus an equivalent definition of random variable is that it is a is a measurable function from Ω into $\mathbb{R}$. It is usually hard to come up with non-measurable functions in practice, so we will not have to be worried about this issue. (Although see the discussion at the end of Example 1.7, and also Exercise 2.3.) △

The mathematical statement $\mu(B) = P(X \in B)$ says that the μ-probability of the set B is the probability that the value of X lies in B. We write μ_X for the distribution of X whenever it is important to highlight the random variable X whose probability distribution is considered.

Example 2.7. Let X_1 be the value of the first roll of a fair die, as in Examples 2.2 and 2.3 above. We illustrate the probability distribution μ_{X_1} on $\mathbb{R}$ by giving some of its values. In each case the probability of a set B comes from the probabilities $\frac{1}{6}$ attached to the values of X_1.

$$\begin{aligned}
\mu_{X_1}\{k\} &= P(X_1 = k) = \tfrac{1}{6} \qquad \text{for } k = 1, 2, 3, 4, 5, 6, \\
\mu_{X_1}([-1, 2]) &= P\{X_1 \in [-1, 2]\} = P\{X_1 = 1 \text{ or } X_1 = 2\} = \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{3}, \\
\mu_{X_1}([100, \infty)) &= P\{X_1 \in [100, \infty)\} = 0, \\
\mu_{X_1}(\mathbb{R}) &= P(X_1 \in \mathbb{R}) = 1.
\end{aligned}$$

△

In a measure-theoretic treatment of random variables the probability distribution μ takes center stage, but we leave this notion in the background. Instead, we focus on the following three ways to describe the probability distribution:

- Every random variable has a *cumulative distribution function*, abbreviated c.d.f.
- A *discrete random variable* has a *probability mass function*, abbreviated p.m.f.
- An *absolutely continuous random variable*[1] has a *probability density function*, abbreviated p.d.f.

We begin with the probability mass function because it is the most accessible.

Discrete random variables and the probability mass function.

Definition 2.8. A random variable X is **discrete** if there exists a finite or countably infinite set $B \subset \mathbb{R}$ such that $P(X \in B) = 1$. △

We say that those values k for which $P(X = k) > 0$ are the *possible values* of the discrete random variable X.

As for functions in general, the *range* of a random variable X is the set of all its values: the range of X is the set $\{X(\omega) : \omega \in \Omega\}$. In particular, if the range of

[1] Absolutely continuous random variables are often called *continuous* in introductory probability textbooks. The term absolutely continuous is technically more accurate, as we will explain later.

the random variable X is finite or countably infinite, then X is a discrete random variable.

Example 2.9. We revisit the examples above to determine which random variables are discrete. Let X_1, X_2, S, N and R be as in Examples 2.3 and 2.4. The next table exhibits the ranges of these random variables and gives an example of a countable set B such that $P(X \in B) = 1$, if such a set exists.

random variable	range	countable set B such that $P(X \in B) = 1$
X_1, X_2	$\{1, \dots, 6\}$	$\{1, \dots, 6\}$
S	$\{2, 3, \dots, 12\}$	$\{2, 3, \dots, 12\}$
Z	$\{0,1,2\}$	$\{0,1,2\}$
N	$\mathbb{Z}_{>0} \cup \{\infty\}$	$\mathbb{Z}_{>0}$
R	the interval $[0, r_0]$	**Does not exist.**

X_1, X_2, S and Z have a finite range, N has a countably infinite range, and R has an uncountable range.

We have $P(N \in \mathbb{Z}_{>0}) = 1$ by Example 1.8. We show that $P(R = r) = 0$ for any particular value r, from which the nonexistence claim in the table follows. For $r \in [0, r_0]$ the event $\{R = r\}$ is the circle of radius r:

$$\{R = r\} = \{(x, y) : x^2 + y^2 = r^2\}.$$

This circle (not the disk inside the circle) has area zero, and so

$$P(R = r) = \frac{\text{area of circle of radius } r}{\pi r_0^2} = \frac{0}{\pi r_0^2} = 0.$$

The conclusion is that X_1, X_2, S and N are discrete random variables, while R is not. △

The notion of a discrete random variable is not defined in terms of the range of the random variable because of examples of the following type where the range is uncountable but all the probability is captured by a discrete set.

Example 2.10. Let the sample space $\Omega = \mathbb{R}$, the real line. Define the probability measure P on Ω by setting $P\{k\} = \frac{1}{3}$ for $k = 1, 2, 3$, and $P(A) = 0$ for all sets A that are disjoint from $\{1, 2, 3\}$. In a single formula we can write

$$P(A) = \tfrac{1}{3} I_A(1) + \tfrac{1}{3} I_A(2) + \tfrac{1}{3} I_A(3)$$

which says that the set A gets a probability of $\frac{1}{3}$ for each of the points $1, 2, 3$ it contains.

Let Y be the identity function: $Y(\omega) = \omega$ for all real ω. Then Y satisfies $P(Y \in \{1, 2, 3\}) = 1$ while the range of Y is the entire real line. Thus Y satisfies the definition of a discrete random variable with possible values $\{1, 2, 3\}$. △

When the random variable X is discrete, a full description of all the probabilities $P(X \in B)$ can be deduced from knowing the probabilities of the possible values. These probabilities are described by the probability mass function.

Definition 2.11. The **probability mass function** (p.m.f.) of a discrete random variable X is the function p (or p_X) defined by

$$p(k) = P(X = k)$$

for the possible values k of X. (In some cases it is convenient to extend the function p for other values as well, we can use the same definition even if $P(X = k) = 0$.) △

Probabilities of events involving X come by summing values of the probability mass function: for any subset $B \subseteq \mathbb{R}$,

$$P(X \in B) = \sum_{k \in B} P(X = k) = \sum_{k \in B} p_X(k), \tag{2.7}$$

where the sum is over the possible values k of X that lie in B.

Example 2.12. (Continuation of Example 2.2.) Here are the probability mass functions of the first die, the sum of the dice, and the number of rolls that give one or two.

k	1	2	3	4	5	6
$p_{X_1}(k) = P(X_1 = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

k	2	3	4	5	6	7	8	9	10	11	12
$p_S(k) = P(S = k)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

k	0	1	2
$p_Z(k) = P(Z = k)$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

The values above are computed from the probabilities of individual sample points. The case $p_S(8) = P(S = 8)$ was done in (2.2). The value $p_Z(1)$ comes from

$$\begin{aligned} p_Z(1) = P(Z = 1) = \frac{1}{36}\#\big[&\{(i,j) : i \in \{1,2\}, j \in \{3,4,5,6\}\} \\ &\cup \{(i,j) : i \in \{3,4,5,6\}, j \in \{1,2\}\}\big] \\ = \frac{2\cdot 4 + 4 \cdot 2}{36} = \tfrac{4}{9}. \end{aligned}$$

Probabilities of events are obtained by summing values of the probability mass function. For example,

$$P(2 \le S \le 5) = p_S(2) + p_S(3) + p_S(4) + p_S(5) = \tfrac{1}{36} + \tfrac{2}{36} + \tfrac{3}{36} + \tfrac{4}{36} = \tfrac{10}{36}.$$

△

Example 2.13. (Continuation of Example 2.3.) We derive the probability mass function of N, the number of rolls needed for the first six. The relevant calculation was already done in Example 1.8. We only need to re-interpret the answer from there:

$$p_N(k) = P(N = k) = \left(\tfrac{5}{6}\right)^{k-1}\tfrac{1}{6} \qquad \text{for} \quad k \in \mathbb{Z}_{>0}.$$

△

Since probabilities are nonnegative and total probability is one, every probability mass function p with finite or countably infinite domain D must satisfy the properties

$$p(k) \geq 0 \ \forall k \in D \qquad \text{and} \qquad \sum_{k \in D} p(k) = 1. \tag{2.8}$$

Conversely, if a function p defined on a finite or countably infinite domain D satisfies (2.8), then p is the probability mass function of a random variable. This is because we can define a discrete probability space $(\Omega, \mathcal{F}, P)$ by setting $\Omega = D$, $\mathcal{F} = 2^D$ (power set), and $P(A) = \sum_{k \in A} p(k)$ for all subsets $A \subset D$. The random variable X we want is the identity function: $X(k) = k$ for all $k \in \Omega$. Then from the definitions, $P(X = k) = P\{k\} = p(k)$, which demonstrates that the p we started with is the probability mass function of X.

Example 2.14. An indicator random variable is always discrete, as it can only take the values 0 and 1. If B is an event then the corresponding indicator variable I_B has a probability mass function given by

$$p(1) = P(I_B = 1) = P(B), \qquad p(0) = p(I_B = 0) = P(B^c) = 1 - P(B).$$

△

Certain probability distributions come up over and over in examples and applications, and hence they have special names.

Definition 2.15. Let $0 \leq p \leq 1$. We say that a random variable X has Bernoulli distribution with parameter p if X is a discrete random variable with probability mass function

$$p_X(1) = p, \qquad p_X(0) = 1 - p.$$

We abbreviate this as $X \sim \text{Ber}(p)$.

The distribution of an indicator random variable I_B is always Bernoulli, its parameter is $P(B)$.

Cumulative distribution function.

Definition 2.16. Let X be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The **cumulative distribution function** (c.d.f.) of X is defined by

$$F(s) = P(X \leq s) \quad \text{for all } s \in \mathbb{R}. \tag{2.9}$$

△

The convention that the inequality is $\leq$ in definition (2.9) is important. The cumulative distribution function gives probabilities of left-open right-closed intervals of the form $(a, b]$:

$$P\{X \in (a, b]\} = P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

To illustrate, we derive the cumulative distribution functions of random variables Z, N and R from the examples above.

For the next example, note that the connection between the probability mass function and the cumulative distribution function of a discrete random variable is

$$F(s) = P(X \le s) = \sum_{k: k \le s} P(X = k), \tag{2.10}$$

where the sum extends over those possible values k of X that are less than or equal to s.

Example 2.17. (Continuation of Example 2.12.) This calculation goes case by case.

$$\begin{aligned} s < 0: &\quad F_Z(s) = P(Z \le s) = 0. \\ 0 \le s < 1: &\quad F_Z(s) = P(Z \le s) = P(Z = 0) = \tfrac{4}{9}. \\ 1 \le s < 2: &\quad F_Z(s) = P(Z \le s) = P(Z = 0) + P(Z = 1) = \tfrac{8}{9}. \\ s \ge 2: &\quad F_Z(s) = P(Z \le s) = P(Z = 0) + P(Z = 1) + P(Z = 2) = 1. \end{aligned}$$

Collecting the cases yields the function

$$F_Z(s) = \begin{cases} 0, & s < 0 \\ \frac{4}{9}, & 0 \le s < 1 \\ \frac{8}{9}, & 1 \le s < 2 \\ 1, & s \ge 2. \end{cases} \tag{2.11}$$

△

Example 2.18. (Continuation of Example 2.13.) Since $N \ge 1$, we can first record $F_N(s) = 0$ for $s < 1$. Then for integers $n \ge 1$ and $s \in [n, n+1)$,

$$\begin{aligned} F_N(s) &= P(N \le s) = P(N \le n) = 1 - P(N > n) = 1 - \sum_{k=n+1}^{\infty} P(N = k) \\ &= 1 - \sum_{k=n+1}^{\infty} \left(\tfrac{5}{6}\right)^{k-1} \tfrac{1}{6} = 1 - \left(\tfrac{5}{6}\right)^{n}. \end{aligned}$$

We switched to the complement above because summing up the tail of the series is simpler than summing up the initial segment. We can summarize the function as follows:

$$F_N(s) = \begin{cases} 0, & s < 1 \\ 1 - \left(\frac{5}{6}\right)^n, & \text{for } n \le s < n+1 \text{ and } n \in \mathbb{Z}_{\ge 1}. \end{cases} \tag{2.12}$$

A less transparent but more compact expression is

$$F_N(s) = 1 - \left(\tfrac{5}{6}\right)^{\lfloor s \rfloor \vee 0}$$

where we used two standard notations: the floor function $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x:

$$\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \le x\} \qquad \text{for } \ x \in \mathbb{R}, \tag{2.13}$$

and $\vee$ is an alternative notation for the maximum of two quantities: $x \vee y = \max\{x, y\} =$ the larger of x and y. △

Example 2.19. (Continuation of Example 2.4) Pick a uniform random point from the disk of radius r_0 centered at the origin. The sample space is $\Omega = \{\omega = (x, y) : x^2 + y^2 \le r_0^2\}$ and probability P is defined by

$$P(A) = \frac{\text{area of } A}{\text{area of } \Omega}$$

for events $A \subset \Omega$. The random variable

$$R(x, y) = \sqrt{x^2 + y^2}$$

gives the distance from the origin to the sample point $\omega = (x, y) \in \Omega$.

To find the cumulative distribution function F_R of R, begin with the obvious values: since $0 \le R(\omega) \le r_0$ for all $\omega \in \Omega$, we have $F_R(s) = P(R \le s) = 0$ for $s < 0$ and $F_R(s) = P(R \le s) = 1$ for $s \ge r_0$. For $0 \le s < r_0$,

$$\begin{aligned} F_R(s) = P(R \le s) &= P\{\text{the chosen point lies in the disk of radius } s\} \\ &= \frac{\text{area of disk of radius } s}{\text{area of disk of radius } r_0} = \frac{\pi s^2}{\pi r_0^2} = \frac{s^2}{r_0^2}. \end{aligned}$$

The cumulative distribution function is then

$$F_R(s) = \begin{cases} 0, & s < 0 \\ s^2/r_0^2, & 0 \le s < r_0 \\ 1, & s \ge r_0. \end{cases} \tag{2.14}$$

△

We develop some general properties of the cumulative distribution function. The first theorem gives precise conditions for when a function is a cumulative distribution function. We cannot fully prove the sufficiency of these conditions, but we can explain what is involved in the proof. The second theorem shows how to find probabilities $P(X < s)$ and $P(X = s)$ from the cumulative distribution function. For the statements below, recall the meaning of one-sided limits $F(x\pm)$ from Appendix C.

Theorem 2.20.

(a) *Let $F : \mathbb{R} \to [0, 1]$ be the cumulative distribution function of a random variable X. Then F has the following properties.*

  (i) *Monotonicity: if $s < t$ then $F(s) \le F(t)$.*

  (ii) *Right continuity: $F(t) = F(t+)$ for each $t \in \mathbb{R}$.*

  (iii) $\lim_{t\to-\infty} F(t) = 0$ *and* $\lim_{t\to\infty} F(t) = 1$.

(b) *Conversely, if a function $F : \mathbb{R} \to [0, 1]$ has properties* (i)–(iii) *above, then F is the cumulative distribution function of some random variable.*

Proof. Part (a). Let F be the cumulative distribution function of X.

(i) $s < t$ implies $F(s) = P(X \le s) \le P(X \le t) = F(t)$ by monotonicity of probability.

(ii) By Lemma C.3 from Appendix C, $F(t) = F(t+)$ follows from showing that $F(s_n) \to F(t)$ for each strictly decreasing sequence $s_n \to t$. Fix such a sequence. We claim that then

$$\{X \le s_n\} \searrow \{X \le t\}. \tag{2.15}$$

To verify (2.15), we need to check two things: that the events $\{X \le s_n\}$ are nested nonincreasing, in other words that

$$\{X \le s_n\} \supset \{X \le s_{n+1}\}, \tag{2.16}$$

and then

$$\{X \le t\} = \bigcap_{n\ge1} \{X \le s_n\}. \tag{2.17}$$

Monotonicity (2.16) is clear from $s_n > s_{n+1}$.

To prove set equality (2.17), we prove $\subset$ and $\supset$. The immediate direction is $\subset$. Since $t < s_n$ for all n, we have $\{X \le t\} \subset \{X \le s_n\}$ for all n, and hence $\{X \le t\} \subset \bigcap_n \{X \le s_n\}$.

For the other direction, assume $\omega \in \bigcap_n \{X \le s_n\}$. This means that the inequality

$$X(\omega) \le s_n$$

holds for all n. Since $s_n \to t$, we can let $n \to \infty$ on the right to get $X(\omega) \le t$. This says $\omega \in \{X \le t\}$. Thus $\bigcap_n \{X \le s_n\} \subset \{X \le t\}$. We have now verified (2.17).

Since both (2.16) and (2.17) hold, (2.15) is valid. By (2.15) and Theorem 1.25 (continuity of probability),

$$F(t) = P(X \le t) = \lim_{n\to\infty} P(X \le s_n) = \lim_{n\to\infty} F(s_n).$$

Right continuity $F(t) = F(t+)$ has been proved.

(iii) To show $\lim_{t\to-\infty} F(t) = 0$, we take an arbitrary $\varepsilon > 0$ and show the existence of s_0 such that $s \le s_0$ implies $F(s) \le \varepsilon$.

Since X is real-valued,

$$\{X \le -n\} \searrow \varnothing. \tag{2.18}$$

Here is the argument. Since $X \le -n-1$ implies $X \le -n$, the events $\{X \le -n\}$ are nested nonincreasing. Let $\omega \in \Omega$ be arbitrary. Then $X(\omega)$ is some real number, and we can take a large enough integer m such that $-m < X(\omega)$. But then $\omega \notin \{X \le -m\}$, and consequently $\omega \notin \bigcap_n \{X \le -n\}$. Thus no sample point lies in $\bigcap_n \{X \le -n\}$.

Again by (2.18) and Theorem 1.25, $F(-n) = P(X \le -n) \to P(\varnothing) = 0$. Given $\varepsilon > 0$, pick n so that $P(X \le -n) \le \varepsilon$. Then $s \le -n$ implies $0 \le F(s) \le F(-n) \le \varepsilon$. This proves $\lim_{t\to-\infty} F(t) = 0$. The remaining limit $\lim_{t\to\infty} F(t) = 1$ is proved similarly and is left as Exercise 2.13.

Part (b). Let F be a function with properties (i)–(iii). We need to show that there exists a random variable X on some probability space $(\Omega, \mathcal{F}, P)$ such that F is the cumulative distribution function of X. This is done as follows. Take $\Omega = \mathbb{R}$, let $\mathcal{F}$ be the Borel σ-algebra of $\mathbb{R}$, and define P on $\mathcal{F}$ by the condition $P((a,b]) = F(b) - F(a)\ \forall -\infty \le a < b \le \infty$. Conditions (i)–(iii) on F and some measure theory show that this defines a unique Borel measure P. Define $X(\omega) = \omega$, the identity function on $\mathbb{R}$. Then by definition, $P(X \le t) = P\{\omega : \omega \le t\} = P((-\infty, t]) = F(t)$. □

Right continuity of F goes together with $\le$ in the definition $F(s) = P(X \le s)$. Probability $P(X < s)$ is a left-continuous function of s (Exercise 2.20).

Theorem 2.21. *Let X be a random variable with cumulative distribution function F. Then for any $s \in \mathbb{R}$ we have these identities:*

$$P(X < s) = F(s-) \tag{2.19}$$

and

$$P(X = s) = F(s) - F(s-). \tag{2.20}$$

Proof. By Lemma C.3 in Appendix C, (2.19) follows from showing that

$$F(s_n) \to P(X < s) \quad \text{as } n \to \infty \tag{2.21}$$

for any strictly increasing sequence s_n that approaches s from the left, that is, $s_n < s_{n+1} < s$ and $s_n \to s$. The proof of (2.21) is based on Theorem 1.25 applied to this set limit:

$$\{X \le s_n\} \nearrow \{X < s\} \quad \text{as } n \to \infty. \tag{2.22}$$

To verify (2.22), we need to check two things: that the events $\{X \le s_n\}$ are nested nondecreasing, in other words that

$$\{X \le s_n\} \subset \{X \le s_{n+1}\}, \tag{2.23}$$

and then

$$\{X < s\} = \bigcup_{n \ge 1} \{X \le s_n\}. \tag{2.24}$$

Monotonicity (2.23) is clear from $s_n < s_{n+1}$.

To prove a set equality such as (2.24), we prove $\subset$ and $\supset$.

The immediate direction is $\supset$. We explain it step by step to develop routine with these types of arguments. If $X(\omega) \le s_n$ for some n, then also $X(\omega) < s$. This says that

$$\omega \in \{X \le s_n\} \quad \text{implies} \quad \omega \in \{X < s\}$$

which is equivalent to the subset relation

$$\{X \le s_n\} \subset \{X < s\}.$$

Since this is true for all $n \ge 1$, the union of the sets on the left lie inside the set on the right:

$$\bigcup_{n \ge 1} \{X \le s_n\} \subset \{X < s\}.$$

This proves one half of (2.24).

For the other direction we assume that $\omega \in \{X < s\}$ and argue that ω also lies in the union on the right-hand side of (2.24). Since s_n approaches s from the left and $X(\omega) < s$, then $X(\omega) < s_m < s$ for some m. This implies that $\omega \in \{X \le s_m\}$, and thereby also $\omega \in \bigcup_{n \ge 1}\{X \le s_n\}$. We have shown that

$$\{X < s\} \subset \bigcup_{n\ge 1}\{X \le s_n\}$$

and completed the proof of (2.24).

Now that both (2.23) and (2.24) have been verified, we have established (2.22). By Theorem 1.25,

$$\lim_{n\to\infty} F(s_n) = \lim_{n\to\infty} P(X \le s_n) = P(X < s).$$

Since the sequence s_n was an arbitrary strictly increasing sequence converging to s, Lemma C.3 establishes $F(s-) = P(X < s)$.

From additivity of probability we verify (2.20):

$$P(X = s) = P(X \le s) - P(X < s) = F(s) - F(s-). \qquad \square$$

Since a cumulative distribution function F is always right-continuous, continuity at s is equivalent to $F(s) = F(s-)$. Thus (2.20) gives us the following statement about the point probabilities of a random variable X with cumulative distribution function F.

Fact 2.22. Let X be a random variable. Then for any $s \in \mathbb{R}$ $P(X = s) = 0$ if and only if F is continuous at s. If F is discontinuous at s, then $P(X = s) =$ the magnitude of the jump in F at point s.

The following theorem allows us to identify discrete random variable from their cumulative distribution function. Its proof is the content of Exercise 2.15. (Exercise 2.16 shows that not all discrete random variables can be identified using this theorem.)

Theorem 2.23. *Let X be a random variable. Suppose that there is a finite or infinite increasing real sequence $s_1 < s_2 < \dots < s_n$, $s_1 < s_2 < \dots$, $\dots < s_{-1} < s_0$, or $\dots < s_{-1} < s_0 < s_1 < \dots$ so that $F(s_k-) < F(s_k)$ and $F(s_k) = F(s_{k+1}-)$ for all k where s_k and s_k, s_{k+1} are defined. Then X is a discrete random variable. The possible values of X are the numbers s_k, and $P(X = s_k) = F(s_k) - F(s_k-)$.*

Probability density function.

Definition 2.24. Let X be a random variable on $(\Omega, \mathcal{F}, P)$. If a function f on $\mathbb{R}$ satisfies $f(x) \ge 0$ for all x and

$$P(X \le b) = \int_{-\infty}^{b} f(x)\,dx \tag{2.25}$$

for all real values b, then f is the **probability density function** (p.d.f.) of X. When X has a density function, we call X an **absolutely continuous** random variable. △

The term *absolutely continuous* is used because this is the property of the cumulative distribution function of X that is both necessary and sufficient for the existence of a density function. (See Section 2.5.) An important technical fact is that if f satisfies Definition 2.24 then

$$P(X \in B) = \int_B f(x)\,dx \tag{2.26}$$

for *any* Borel subset B of the real line. This includes all sets that arise in practice. This is not an obvious consequence of Definition 2.24. See Section 2.5♣ for additional explanation.

Common examples of the set B in (2.26) include intervals, bounded or unbounded, and collections of intervals. These give identities such as

$$P(a \le X \le b) = \int_a^b f(x)\,dx \quad \text{and} \quad P(X > a) = \int_a^\infty f(x)\,dx \tag{2.27}$$

for any real $a \le b$. Furthermore, if random variable X has density function f then point values have probability zero:

$$P(X = c) = \int_c^c f(x)\,dx = 0 \qquad \text{for any real } c. \tag{2.28}$$

This has two concrete consequences. (i) Absolutely continuous and discrete random variables are completely separate classes of random variables. (ii) Probabilities of intervals for absolutely continuous random variables are not changed by including or excluding endpoints. Thus in (2.27) we have also

$$P(a < X \le b) = P(a < X < b) = \int_a^b f(x)\,dx \quad \text{and} \quad P(X \ge a) = \int_a^\infty f(x)\,dx.$$

Equation (2.25) gives the connection between the probability density function and the cumulative distribution function:

$$F(x) = \int_{-\infty}^x f(s)\,ds \qquad \text{for all } x \in \mathbb{R}. \tag{2.29}$$

Differentiating both sides gives the equation

$$F'(x) = f(x) \qquad \text{for all } x \in \mathbb{R} \text{ at which } f \text{ is continuous.} \tag{2.30}$$

We state the following theorem that gives a sufficient condition on the cumulative distribution function that implies the existence of a probability density function.

Theorem 2.25. *Suppose the cumulative distribution function F of the random variable X is continuous and the derivative $F'(x)$ exists everywhere on the real line, except possibly at countably many points. Then X is an absolutely continuous random variable and $f(x) = F'(x)$ is the density function of X. If F is not differentiable at a point x, then the value $f(x)$ can be set arbitrarily.* ♣

We illustrate the theorem with an earlier example.

Example 2.26. (Continuation of Example 2.19) The cumulative distribution function found in Example 2.19 was

$$F_R(s) = \begin{cases} 0, & s < 0 \\ s^2/r_0^2, & 0 \le s < r_0 \\ 1, & s \ge r_0. \end{cases}$$

This is a continuous function that is differentiable everywhere apart from $s = r_0$ where the slope from the left is $2/r_0$ but the slope from the right is zero. Thus we can take the density function to be

$$f_R(s) = \begin{cases} 2s/r_0^2, & 0 \le s < r_0 \\ 0, & s < 0 \text{ or } s \ge r_0. \end{cases} \tag{2.31}$$

△

Next we state a theorem that characterizes density functions. Then we can appeal to it to write down examples. We state it for piecewise continuous functions (see Definition C.4 in Appendix C) because that is general enough for us and it allows us to avoid technical complications about existence of integrals.

Theorem 2.27. *Let f be a piecewise continuous function on $\mathbb{R}$. Then f is the density function of a random variable if and only if*

$$f(x) \ge 0 \textit{ for all } x \in \mathbb{R} \quad \textit{and} \quad \int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{2.32}$$

Proof idea. Assuming that f is a density function of a random variable X, $f \ge 0$ was already part of Definition 2.24, and the integral condition follows from

$$\int_{-\infty}^{\infty} f(x)\,dx = P(-\infty < X < \infty) = 1.$$

Conversely, if f satisfies (2.32), we define $F(x) = \int_{-\infty}^{x} f(s)\,ds$ for real x. It can be proved that this F satisfies conditions (i)–(iii) of Theorem 2.20, in other words, that F is the cumulative distribution function of some random variable X. Then f is the density function of X. □

Henceforth we know that any f that satisfies (2.32) is a legitimate probability density function. This allows us to write down plenty of examples. Important examples have names. Here is the first one.

Definition 2.28. (Uniform distribution on an interval.) Let $[a, b]$ be a bounded interval on the real line. Random variable X has the *uniform distribution on the interval* $[a, b]$ if X has density function

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{if } x \notin [a, b]. \end{cases} \tag{2.33}$$

Abbreviate this by $X \sim \text{Unif}[a, b]$. △

If $X \sim \text{Unif}[a, b]$ and $[c, d] \subset [a, b]$, then

$$P(c \le X \le d) = \int_c^d \frac{1}{b-a}\,dx = \frac{d-c}{b-a}, \tag{2.34}$$

the ratio of the lengths of the intervals $[c, d]$ and $[a, b]$. In particular, we see that Definition 2.28 generalizes Example 1.9 where the case $[a, b] = [0, 1]$ was considered.

By (2.28) individual points make no difference to any probability calculation with a density function. Hence in Definition 2.28 we can drop one or both endpoints a and b if we so prefer, and define a uniform random variable on the half-open interval $(a, b]$ or on the open interval (a, b). It makes no difference to any probability calculation because in any case $P(X = a \text{ or } X = b) = 0$.

Remark 2.29. The value of a density function at a given point is not a probability. (Note that it can be larger than 1!) However, if the density function is continuous at a given point, then it does provide useful information about probabilities related to the random variable. Let X be a random variable with density function f, and assume that f is continuous at $x = a$. For any $\varepsilon > 0$ we have

$$P(X \in (a, a+\varepsilon)) = \int_a^{a+\varepsilon} f(x)dx.$$

By assumption f is continuous at $x = a$, which means that for small ε the value of $f(x)$ is close to $f(a)$ on $(a, a+\varepsilon)$. Hence $\int_a^{a+\varepsilon} f(x)dx$ is close to $\varepsilon f(a)$ for small ε:

$$P(X \in (a, a+\varepsilon)) \approx \varepsilon f(a).$$

This statement can be made rigorous using limits (see Exercise 2.26):

$$\lim_{\varepsilon \to 0^-} \frac{1}{\varepsilon} P(X \in (a, a+\varepsilon)) = f(a). \tag{2.35}$$

Hence, if the density function f of a random variable X is continuous at a given $a \in \mathbb{R}$ then the value $f(a)$ describes the asymptotic probability of X being in a small interval near a. △

Remark 2.30. The fact that changing a function at a point does not alter the value of an integral creates a uniqueness problem for probability density functions. For example, let $c \in (a, b)$ and define the function

$$\widetilde{f}(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, c) \cup (c, b] \\ 0, & \text{if } x \notin [a, c) \cup (c, b]. \end{cases}$$

In other words, to define $\widetilde{f}$ we changed the function f in (2.33) by setting the value at c to zero. The two functions f and $\widetilde{f}$ are not equal, but over any interval their integrals agree:

$$\int_u^v f(x)\,dx = \int_u^v \widetilde{f}(x)\,dx \qquad \text{for all } -\infty \le u \le v \le \infty.$$

Thus both f and $\widetilde{f}$ can be regarded as legitimate density functions for the Unif$[a, b]$ distribution because they give the same probabilities for every event of the form $P(X \in B)$. We prefer f because it is simpler.

A way around this lack of uniqueness would be to identify named probability distributions of random variables in terms of their cumulative distribution functions because the cumulative distribution function is uniquely defined. However, it is customary to use density functions. The reader just has to keep in mind that it is the probabilities of events $P(X \in B)$ that are of primary importance, and not the point values $f(x)$ of a probability density function. △

Random variables that are neither discrete nor absolutely continuous.

Discrete and absolutely continuous random variables are special types of random variables. These two categories cover the most important random variables, but they do not by any means exhaust the possibilities. Consider this example.

Example 2.31. Fix $a < b$ and define $F : \mathbb{R} \to [0, 1]$ by

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{3} \cdot \frac{x-a}{b-a} & \text{if } a \le x < b, \\ 1 & \text{if } x \ge b. \end{cases}$$

Checking the conditions of Theorem 2.20 confirms that F is the cumulative distribution function of some random variable X. From the formula,

$$P(X = b) = F(b) - F(b-) = \tfrac{2}{3},$$

so X cannot be absolutely continuous. Might X be discrete? F has only one jump which is at b, and so $P(X = x) = 0$ for all $x \ne b$. For any countable set $B = \{x_1, x_2, x_3, \dots\}$,

$$P(X \in B) = \sum_i P(X = x_i) = \begin{cases} \frac{2}{3}, & \text{if } b \in B \\ 0, & \text{if } b \notin B. \end{cases}$$

Thus no countable set B satisfies $P(X \in B) = 1$. Consequently X is not discrete either. △

The example above is a mixture of absolutely continuous and discrete, because it behaves like a uniform random variable on $[a, b)$, but then puts positive probability on the point b. The next word problem illustrates how naturally an example like this arises.

Example 2.32. Suppose your homeowner's insurance has a deductible of 2000 dollars. When the old locust next to your house falls on the roof, the cost of repairs is uniformly distributed between 1000 and 5000 dollars. Let Y denote the cost of the repairs and X the amount you pay, both in units of 1000 dollars. Find the cumulative distribution functions of Y and X.

Since $Y \sim \text{Unif}[1, 5]$, the cumulative distribution function of Y is

$$F_Y(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{x-1}{4} & \text{if } 1 \le x < 5, \\ 1 & \text{if } x \ge 5. \end{cases}$$

When $Y < 2$, $X = Y$ because you have to pay the entire cost of the repairs. But once $Y \ge 2$, you pay only 2 (this is the meaning of a deductible), and then $X = 2$. Thus for $x < 2$, $P(X \le x) = P(Y \le x)$. Since 2 is the upper bound for X, $P(X \le x) = 1$ for all $x \ge 2$. We collect this into a formula:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1, \\ \frac{x-1}{4} & \text{if } 1 \le x < 2, \\ 1 & \text{if } x \ge 2. \end{cases}$$

Let us observe that X has both features of a uniform distribution and a discrete distribution. For $1 < a < b < 2$,

$$P(a < X \le b) = F_X(b) - F_X(a) = \frac{b-a}{4}$$

while

$$P(X = 2) = P(Y \ge 2) = \tfrac{5-2}{4} = \tfrac{3}{4}.$$

This point mass can also be read off from F_X:

$$\begin{aligned} P(X = 2) &= F(2) - F(2-) = F(2) - \lim_{x \to 2-} F(x) \\ &= 1 - \lim_{x \to 2-} \tfrac{x-1}{4} = 1 - \tfrac{1}{4} = \tfrac{3}{4}. \end{aligned}$$

△

Even mixtures of discrete and absolutely continuous random variables do not cover all cases. The *Cantor function* is a famous example of a distribution function that is continuous everywhere on $\mathbb{R}$ and yet no part of it is an integral of a density function.

2.2. Random vectors

When several random variables $X_1, X_2, \dots, X_d$ are defined on the same probability space, we can combine them into a random vector $\mathbf{X} = (X_1, X_2, \dots, X_d)$ whose values are in $\mathbb{R}^d$. The concepts introduced in Section 2.1 generalize naturally to random vectors. Throughout this section, d is a positive integer.

Definition 2.33. Let $(\Omega, \mathcal{F}, P)$ be a probability space and d a positive integer. A d-dimensional **random vector** on Ω is a function $\mathbf{X} : \Omega \to \mathbb{R}^d$.

The probability distribution μ of a d-dimensional random vector $\mathbf{X}$ is a probability measure μ on $\mathbb{R}^d$, defined by

$$\mu(B) = P(\mathbf{X} \in B) \qquad \text{for subsets } B \subset \mathbb{R}^d. \tag{2.36}$$

As before, $\mu(B)$ is defined only for Borel subsets B of $\mathbb{R}^d$. Since a random vector $\mathbf{X} = (X_1, X_2, \dots, X_d)$ has coordinates X_i which are themselves real-valued random variables, the probability distribution of $\mathbf{X}$ is also called the *joint distribution* of the random variables $(X_1, X_2, \dots, X_d)$. In this context, the distributions of the individual coordinates X_i and of the smaller vectors such as (X_1, X_2) (if $d \ge 3$) are called *marginal distributions*.

In the sequel we use boldface symbols for vector quantities, as for $\mathbf{X}$ above, and also for example $\mathbf{k} = (k_1, \dots, k_d)$ and $\mathbf{x} = (x_1, \dots, x_d)$.

Definition 2.34. Suppose the random variables $X_1, X_2, \dots, X_d$ are discrete. Then the **joint probability mass function** of $\mathbf{X} = (X_1, X_2, \dots, X_d)$ is the function

$$p_{\mathbf{X}}(\mathbf{k}) = P(\mathbf{X} = \mathbf{k}) = P(X_1 = k_1, X_2 = k_2, \dots, X_d = k_d)$$

defined for all possible values k_1 of X_1, k_2 of X_2, …, k_d of X_d. $\mathbf{X}$ is a **discrete random vector**. △

Marginal probability mass functions of subsets of the variables $X_1, X_2, \dots, X_d$ are obtained by summing over the possible values of the other random variables. We state two cases in the next theorem. The principle would be the same for any smaller vector of coordinates X_i.

Theorem 2.35.

(a) *Let $p_{\mathbf{X}}$ be the joint probability mass function of $\mathbf{X} = (X_1, \dots, X_d)$. Let $1 \le j \le d$. Then the marginal probability mass function of X_j is given by*

$$p_{X_j}(k) = \sum_{\ell_1,\dots,\ell_{j-1},\ell_{j+1},\dots,\ell_d} p_{\mathbf{X}}(\ell_1, \dots, \ell_{j-1}, k, \ell_{j+1}, \dots, \ell_d), \tag{2.37}$$

where the sum is over the possible values of the other random variables.

(b) *Let $1 \le m < d$. The joint probability mass function of $(X_1, \dots, X_m)$ is obtained from*

$$p_{X_1,\dots,X_m}(k_1, \dots, k_m) = \sum_{\ell_{m+1},\dots,\ell_d} p_{\mathbf{X}}(k_1, \dots, k_m, \ell_{m+1}, \dots, \ell_d)$$

where the sum ranges over all possible values $\ell_{m+1}, \dots, \ell_d$ of the random variables $X_{m+1}, \dots, X_d$.

Proof. We prove part (a). The proof decomposes the event $\{X_j = k\}$ according to the different values of the other random variables $X_1, X_2, \dots, X_{j-1}, X_{j+1}, \dots, X_d$.

$$\begin{aligned} p_{X_j}(k) &= P(X_j = k) \\ &= \sum_{\ell_1,\dots,\ell_{j-1},\ell_{j+1},\dots,\ell_d} P(X_1 = \ell_1, \dots, X_{j-1} = \ell_{j-1}, X_j = k, \\ &\qquad\qquad X_{j+1} = \ell_{j+1}, \dots, X_d = \ell_d) \\ &= \sum_{\ell_1,\dots,\ell_{j-1},\ell_{j+1},\dots,\ell_d} p(\ell_1, \dots, \ell_{j-1}, k, \ell_{j+1}, \dots, \ell_d). \end{aligned}$$

In the sums above, index ℓ_i ranges over all the possible values of X_i, for $i \neq j$. □

Example 2.36. (Continuation of Example 2.3.) Consider $\mathbf{X} = (X_1, X_2, S, Z, N)$ as a random vector on sample space (2.3), defined by equations (2.4) and (2.5). Two particular values of the joint probability mass function are calculated as follows:

$$\begin{aligned} p_{\mathbf{X}}(3, 4, 7, 0, 4) &= P(X_1 = 3, X_2 = 4, S = 7, Z = 0, N = 4) \\ &= P\{\omega : s_1 = 3, s_2 = 4, s_3 \neq 6, s_4 = 6\} = \frac{1 \cdot 1 \cdot 5 \cdot 1}{6^4} = \tfrac{5}{1296} \end{aligned}$$

and

$$p_{\mathbf{X}}(3, 3, 7, 0, 4) = P(X_1 = 3, X_2 = 3, S = 7, Z = 0, N = 4) = 0.$$

The second probability is zero because the event in question is empty: $X_1(\omega) = X_2(\omega) = 3$ force ω to have $s_1 = s_2 = 3$, while $S(\omega) = 7$ forces $s_1 + s_2 = 7$, and no sequence ω can satisfy both conditions simultaneously.

When there are two random variables, the joint probability mass function can be represented by a table. We illustrate with the random vector (Z, X_1).

		X_1					
		1	2	3	4	5	6
Z	0	0	0	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$
	1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$	$\frac{1}{18}$
	2	$\frac{1}{18}$	$\frac{1}{18}$	0	0	0	0

The numbers in the cells of the table are the probabilities of pairs of values. For example,

$$P(Z = 1, X_1 = 2) = P\{\omega : (s_1, s_2) = (2, j),\, j \notin \{1, 2\}\} = \tfrac{4}{36} = \tfrac{1}{9}$$

while

$$P(Z = 2, X_1 = 4) = P(\varnothing) = 0.$$

The row sums of the table give the marginal probability mass function of Z and the column sums give the marginal probability mass function of X_1. For example,

$$\begin{aligned} P(X_1 = 4) &= P(X_1 = 4, Z = 0) + P(X_1 = 4, Z = 1) + P(X_1 = 4, Z = 2) \\ &= \tfrac{1}{9} + \tfrac{1}{18} + 0 = \tfrac{1}{6}. \end{aligned}$$

△

Definition 2.37. The **joint cumulative distribution function** of the random variables $\mathbf{X} = (X_1, \ldots, X_d)$ is

$$\begin{aligned} F(\mathbf{x}) &= F_{X_1,\ldots,X_d}(x_1, \ldots, x_d) \\ &= P(X_1 \le x_1, \ldots, X_d \le x_d) \quad \text{for } \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d. \end{aligned} \tag{2.38}$$

△

Definition 2.38. Random variables $\mathbf{X} = (X_1, \ldots, X_d)$ are **jointly absolutely continuous** if there exists a function $f : \mathbb{R}^d \to [0, \infty)$ such that, for all $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(y_1, \ldots, y_d)\, dy_1 \cdots dy_d. \tag{2.39}$$

When it exists, f is the **joint probability density function** of $\mathbf{X}$. △

Again it follows that if a joint density function f exists, then for any (Borel) subset $B \subset \mathbb{R}^d$,

$$P(\mathbf{X} \in B) = \int_B f(\mathbf{x})\, d\mathbf{x}. \tag{2.40}$$

This point is expanded upon in Section 2.5.♣Above we expressed an integral over a subset B of $\mathbb{R}^d$ with a single integral sign and abbreviated $dx_1 \cdots dx_d$ as $d\mathbf{x}$.

If the joint density function is continuous at $\mathbf{x}$, the value $f(\mathbf{x})$ can be found from (2.39) by d-fold partial differentiation:

$$f(x_1, \dots, x_d) = \frac{\partial^d}{\partial x_1 \cdots \partial x_d} F(x_1, \dots, x_d). \tag{2.41}$$

Joint cumulative distribution functions and density functions can be characterized analogously to their scalar counterparts in Theorems 2.20 and 2.27. We state below the result for the joint density function so that we can refer to it for constructing examples. The term *Borel function* in the theorem is used simply to give an accurate statement. Every function that arises in practice falls under this category. Proof of the "if" part of the theorem requires again the construction of a probability space and a random vector, which we do not address.

Theorem 2.39. *Suppose f is a Borel function on $\mathbb{R}^d$. Then f is the joint density function of a random vector if and only if $f(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in \mathbb{R}^d$ and $\int_{\mathbb{R}^d} f(\mathbf{x})\,d\mathbf{x} = 1$.*

Example 2.40. Let X and Y be two random variables and suppose the vector (X, Y) has a joint density function f. An immediate consequence of (2.40) is that if B is a subset of the plane with zero area, then $P\{(X, Y) \in B\} = 0$. An example of a zero area set is $D = \{(x, y) \in \mathbb{R}^2 : x = y\}$, the diagonal line on the plane. Since $X = Y$ is equivalent to $(X, Y) \in D$, we get the important conclusion that

$$P(X = Y) = 0 \text{ whenever } (X, Y) \text{ has a joint density function } f. \tag{2.42}$$

We can deduce this also by iterated integration:

$$\begin{aligned} P(X = Y) = P\{(X, Y) \in D\} &= \iint_D f(x, y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} \Bigl(\int_x^x f(x, y)\,dy \Bigr)\,dx = \int_{-\infty}^{\infty} 0\,dx = 0. \end{aligned} \tag{2.43}$$

△

By analogy with Theorem 2.35, integrating away variables from a joint density function produces the marginal density function of a smaller set of variables. We state the absolutely continuous analogues of the two cases of Theorem 2.35. Formula (2.44) below says that to find $f_{X_j}(x)$, place x in the jth coordinate inside f, and integrate away the other $d-1$ variables.

Theorem 2.41.

(a) *Let f be the joint density function of $(X_1, \dots, X_d)$. Then each random variable X_j for $1 \le j \le d$ has a marginal density function f_{X_j} that can be obtained by integrating away the other variables from f:*

$$f_{X_j}(x) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-1 \text{ integrals}} f(x_1, \dots, x_{j-1}, x, x_{j+1}, \dots, x_d) \\ dx_1 \dots dx_{j-1}\,dx_{j+1} \dots dx_d. \tag{2.44}$$

(b) *Let $1 \leq m < d$. The joint probability density function of $(X_1, \dots, X_m)$ is obtained from*

$$f_{X_1,\dots,X_m}(x_1, \dots, x_m) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{d-m \text{ integrals}} f(x_1, \dots, x_m, y_{m+1}, \dots, y_d)\, dy_{m+1} \cdots dy_d. \tag{2.45}$$

Proof. We demonstrate the proof for two random variables (X, Y) with joint density function $f_{X,Y}$. For two random variables X and Y the formula is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy. \tag{2.46}$$

By Definition 2.25 of a probability density function we need to show that for any real b,

$$P(X \leq b) = \int_{-\infty}^{b} \left(\int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy \right) dx.$$

This is straightforward since

$$P(X \leq b) = P(-\infty < X \leq b,\ -\infty < Y < \infty) = \int_{-\infty}^{b} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy\, dx,$$

where the final equality is identity (2.40) for the set

$$B = \{(x, y) : -\infty < x \leq b,\ -\infty < y < \infty\}.$$ □

Example 2.42. Let (X, Y) be the coordinates of a uniform random point on the sample space $\Omega = \{(x, y) : x^2 + y^2 \leq r_0^2\}$. In other words, $(X(\omega), Y(\omega)) = (x, y)$ for $\omega = (x, y) \in \Omega$. For any subset $B \subset \mathbb{R}^2$,

$$P((X, Y) \in B) = \frac{\text{area of } B \cap \Omega}{\text{area of } \Omega} = \int_{B \cap \Omega} \frac{1}{\pi r_0^2}\, dx\, dy = \int_B f(x, y)\, dx\, dy$$

where the joint density function f is defined by

$$f(x, y) = \begin{cases} \frac{1}{\pi r_0^2}, & (x, y) \in \Omega \\ 0 & (x, y) \notin \Omega. \end{cases}$$

The marginal density function of X is

$$f_X(x) = \int\limits_{-\infty}^{\infty} f(x, y)\, dy = \frac{1}{\pi r_0^2} \int\limits_{-\sqrt{r_0^2 - x^2}}^{\sqrt{r_0^2 - x^2}} dy = \frac{2}{\pi r_0^2} \sqrt{r_0^2 - x^2}$$

for $-r_0 \leq x \leq r_0$. Since the situation for Y is the same as for X, we also have

$$f_Y(y) = \frac{2}{\pi r_0^2} \sqrt{r_0^2 - y^2} \qquad \text{for } -r_0 \leq y \leq r_0.$$

△

Example 2.42 above is a special case of the following construction which defines a uniform random point on any subset Ω of a Euclidean space $\mathbb{R}^d$, as long as Ω has a finite volume. Here we use *volume* in a sense that naturally generalizes two-dimensional physical area and three-dimensional physical volume. For a d-dimensional rectangle

$$B = \prod_{i=1}^{d}[a_i, b_i] = \{x \in \mathbb{R}^d : a_i \le x_i \le b_i \text{ for } i = 1, \dots, d\}$$

the volume is

$$\operatorname{vol}(B) = \prod_{i=1}^{d}(b_i - a_i).$$

Definition 2.43. Let Ω be a subset of d-dimensional Euclidean space $\mathbb{R}^d$ with finite volume. Then the random point $\mathbf{X}$ is **uniformly distributed on** Ω if its joint density function is

$$f(\mathbf{x}) = \begin{cases} \dfrac{1}{\operatorname{vol}(\Omega)} & \text{if } \mathbf{x} \in \Omega \\ 0 & \text{if } \mathbf{x} \notin \Omega. \end{cases} \tag{2.47}$$

△

Remark 2.44. (Nonexistence of the joint density function.) Theorem 2.41 states that if a random vector $(X_1, \dots, X_d)$ has a joint density function, then the individual coordinates X_j have density functions. The converse implication is not true. This example illustrates.

Let X be a random variable with density function f_X and define Y by $Y(\omega) = X(\omega)$ for $\omega \in \Omega$. In other words, Y is an exact copy of X, defined on the same sample space Ω as X. Y has the same density function as X. But the pair (X, Y) cannot have a joint density function because now $P(X = Y) = 1$, in contradiction with (2.42).

The geometric explanation is that the values of (X, Y) lie on the diagonal which has zero area, and consequently a joint density function cannot exist. △

2.3. Notions of equality

On the most basic level, two functions f and g are equal if their domains are the same and if $f(x) = g(x)$ for all x in their common domain. In this case we write $f = g$ and say that functions f and g are *pointwise equal*. Similarly, two random variables X and Y defined on the same sample space Ω are pointwise equal if $X(\omega) = Y(\omega)$ for all $\omega \in \Omega$.

Pointwise equality is not ideally suited for random variables because it ignores the probability measure. Below we introduce two notions of equality of random variables, These notions depend both on the random variables themselves and on the probability measures on the sample spaces on which they are defined. The second one, equality in distribution, is used in the definition of exchangeable random variables in Section 3.4.

Almost sure equality.

Definition 2.45. Let X and Y be random variables defined on $(\Omega, \mathcal{F}, P)$. Then X and Y are **equal almost surely** if $P(X = Y) = 1$. This is abbreviated by $X = Y$ a.s. $\triangle$

Almost sure equality is also expressed by saying $X = Y$ *with probability one*, abbreviated $X = Y$ w.p.1. Below is a discrete and an absolutely continuous example of almost sure equality $X = Y$ where pointwise equality fails.

Example 2.46. Let $\Omega = \{1, 2, 3\}$ with probability measure $P\{1\} = P\{2\} = \frac{1}{2}$ and $P\{3\} = 0$. Define random variables X and Y on Ω by

$$X(1) = Y(1) = 1, \;\; X(2) = Y(2) = 2, \;\; X(3) = 3 \;\; \text{and} \;\; Y(3) = 0.$$

Then $P(X = Y) = P\{1, 2\} = 1$. $\triangle$

Example 2.47. Let $\Omega = [0, 1]$ and P the probability measure of Example 1.9 that satisfies $P(I) =$ length of I for intervals $I \subset \Omega$. Define random variables X and Y on Ω by $X(\omega) = 0$ for all $\omega \in \Omega$ while $Y(\frac{1}{2}) = \frac{1}{2}$ and $Y(\omega) = 0$ for all $\omega \in \Omega \setminus \{\frac{1}{2}\}$. Then $P(X \neq Y) = P\{\frac{1}{2}\} = 0$. $\triangle$

Equality in distribution.

Definition 2.48. Random variables X and Y are **equal in distribution** if $P(X \in B) = P(Y \in B)$ for all Borel subsets B of $\mathbb{R}$. This is abbreviated by $X \overset{d}{=} Y$. $\triangle$

Thus X and Y are equal in distribution if and only if their probability distributions μ_X and μ_Y on $\mathbb{R}$ coincide. In contrast with Definition 2.45, equality in distribution does not require that X and Y be defined on the same sample space.

Equality in distribution $X \overset{d}{=} Y$ can be established by verifying that X and Y have the same cumulative distribution functions, or the same probability mass functions when X and Y are discrete, or the same probability density functions when X and Y are absolutely continuous. This follows because probability distributions are uniquely represented by cumulative distribution functions always, by probability mass functions in the discrete case, and by probability density functions in the absolutely continuous case.

We give one discrete and one absolutely continuous example.

Example 2.49. Flip a fair coin repeatedly until the first heads. Let X be the number of flips needed. Roll a fair die repeatedly until the die gives an even number. Let Y be the number of rolls needed. Then both X and Y have the same probability mass function: $P(X = k) = P(Y = k) = 2^{-k}$ for $k \in \mathbb{Z}_{>0}$. In particular, $X \overset{d}{=} Y$. $\triangle$

Example 2.50. Fix a real $r_0 > 0$.

Let $\Omega = [0, 1]$ and P the probability measure of Example 1.9 that satisfies $P(I) =$ length of I for intervals $I \subset \Omega$. Define random variable Z on Ω by $Z(\omega) = r_0\sqrt{\omega}$. The range of Z is $[0, r_0]$. For $0 \leq s \leq r_0$ and $\omega \in [0, 1]$, $Z(\omega) \leq s$ if and only if $\omega \leq (s/r_0)^2$. Hence for $0 \leq s \leq r_0$,

$$F_Z(s) = P(Z \leq s) = P\{[0, \tfrac{s^2}{r_0^2}]\} = \tfrac{s^2}{r_0^2}.$$

Thus the full cumulative distribution function of Z is given by

$$F_Z(s) = \begin{cases} 0, & s < 0 \\ s^2/r_0^2, & 0 \le s < r_0 \\ 1, & s \ge r_0. \end{cases}$$

Let R be the random variable of Example 2.4, namely, the distance from the center to a uniform random point on a disk of radius r_0. Its cumulative distribution function F_R was found in Example 2.19. Comparison of F_Z with (2.14) shows that $F_R = F_Z$. Thus $R \overset{d}{=} Z$. △

An immediate theorem is that almost sure equality is stronger than equality in distribution.

Theorem 2.51. *Suppose X and Y are random variables on the same probability space $(\Omega, \mathcal{F}, P)$. Then $P(X = Y) = 1$ implies $X \overset{d}{=} Y$.*

Proof. Let B be a Borel subset of $\mathbb{R}$. Then

$$\begin{aligned} P(X \in B) = P(X \in B, X = Y) + P(X \in B, X \ne Y) = P(Y \in B, X = Y) + 0 \\ = P(Y \in B, X = Y) + P(Y \in B, X \ne Y) = P(Y \in B). \end{aligned}$$

□

Random vectors.

These notions of equality extend to random vectors in the obvious way.

Definition 2.52. Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors defined on $(\Omega, \mathcal{F}, P)$. Then $\mathbf{X}$ and $\mathbf{Y}$ are **equal almost surely** if $P(\mathbf{X} = \mathbf{Y}) = 1$. This is abbreviated by $\mathbf{X} = \mathbf{Y}$ a.s. △

Definition 2.53. Let $\mathbf{X}$ and $\mathbf{Y}$ be $\mathbb{R}^d$-valued random vectors. Then $\mathbf{X}$ and $\mathbf{Y}$ are **equal in distribution** if $P(\mathbf{X} \in B) = P(\mathbf{Y} \in B)$ for all Borel subsets B of $\mathbb{R}^d$. This is abbreviated by $\mathbf{X} \overset{d}{=} \mathbf{Y}$. △

Theorem 2.54. *Let $\mathbf{X}$ and $\mathbf{Y}$ be $\mathbb{R}^d$-valued random vectors defined on the same probability space $(\Omega, \mathcal{F}, P)$. Then $P(\mathbf{X} = \mathbf{Y}) = 1$ implies $\mathbf{X} \overset{d}{=} \mathbf{Y}$.*

2.4. Functions of random variables and vectors

As we have seen, the distribution of a random variable $X : \Omega \to \mathbb{R}$ can be described various ways: using the cumulative distribution function, the probability mass function, or the probability density function. If $g : \mathbb{R} \to \mathbb{R}$ is a measurable function then $Y = g(X)$ is also a random variable, and its distribution is completely determined by g and the distribution of X:

$$P(Y \in B) = P(g(X) \in B) = P(X \in g^{-1}(B)).$$

(Here $g^{-1}(B) = \{\omega : g(\omega) \in B\}$ is the inverse image of B via g.) Hence it is natural to ask how we can express cumulative distribution function, the probability mass function, or the probability density function of $Y = g(X)$ using information about

X and g. The same question can be asked about functions of vector valued random variables.

The cumulative distribution function of $Y = g(X)$ can usually be identified if there is a 'simple' solution for the inequality $g(a) \le x$ in x. Let $A_x = \{a : g(a) \le x\}$ be the solution set of this inequality for a given $x \in \mathbb{R}$. Then

$$F_Y(y) = P(Y \le y) = P(g(X) \le y) = P(X \in A_y).$$

If the solution set A_y is an interval, or the union of a couple of intervals then $P(X \in A_y)$ can be expressed using the cumulative distribution function F_X.

Example 2.55. Express the cumulative distribution function of $Y = (X-2)^2$ with the cumulative distribution function of X.

Since $Y \ge 0$, we have $F_Y(y) = 0$ for $y < 0$. We also have

$$F_Y(0) = P(Y \le 0) = P((X-2)^2 \le 0) = P(X = 2) = F_X(2) - F_X(2-)$$

where we used (2.20). Finally, if $y > 0$ then the solution set of $(x-2)^2 \le y$ is the interval $[2-\sqrt{y}, 2+\sqrt{y}]$, hence in this case

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P((X-2)^2 \le y) = P(2-\sqrt{y} \le X \le 2+\sqrt{y}) \\ &= F_X(2+\sqrt{y}) - F_X(2-\sqrt{y}-). \end{aligned}$$

If F_X happens to be a continuous function then $F_X(a-) = F_X(a)$ for all a, and then the previous expressions simplify a bit. △

Recall the definition of a discrete random variable from Definition 2.8. If X is a discrete random variable, then $Y = g(X)$ is discrete as well: if B is a finite or countably infinite set with $P(X \in B) = 1$ then $g(B) = \{g(a) : a \in B\}$ is also finite or countably infinite and $P(Y \in g(B) = 1)$. If $g : \mathbb{R} \to \mathbb{R}$ has a finite or countable infinite range then $g(X)$ is discrete for any choice of X, since $g(X)$ is always an element of the range of g. In both of these situations the probability mass function of $g(X)$ can sometimes be computed directly.

Example 2.56. Suppose that X is an absolutely continuous random variable with probability density function given by

$$f(x) = \begin{cases} \frac{1}{x^2}, & \text{if } x \ge 1, \\ 0, & \text{if } x < 0. \end{cases}$$

$f(x) = \frac{1}{x^2} I(x \ge 1)$. Let $Y = \lfloor X \rfloor$, where $\lfloor x \rfloor$ is the integer part of x (the largest integer which is at most as large as x). Show that Y is discrete and find its probability mass function.

The range of the function $x \to \lfloor x \rfloor$ is $\mathbb{Z}$, which is countably infinite. Hence $Y = \lfloor X \rfloor$ is discrete. For a given integer k we have $\lfloor x \rfloor = k$ if and only if $k \le x < x+1$. Hence

$$P(Y = k) = P(\lfloor X \rfloor = k) = P(k \le X < k+1) = \int_k^{k+1} f(x)dx.$$

Evaluating the integral we get $P(Y = k) = 0$ for $k \le 0$ and $P(Y = k) = \frac{1}{k} - \frac{1}{k+1} = \frac{1}{k(k+1)}$ for $k \ge 1$. △

If X is absolutely continuous with probability density function f_X and g is 'nice' enough then $Y = g(X)$ will be absolutely continuous as well. E.g. if g is strictly increasing and differentiable then this is true, but this is not a necessary condition. Often the best approach is to derive an expression for the cumulative distribution function of Y in terms of F_X, and then to differentiate it to get the probability density function of Y.

Example 2.57. Suppose that X is an absolutely continuous random variable with a continuous density f_X. Show that $Y = (X-2)^2$ is also absolutely continuous, and derive its probability density function.

In Example 2.55 we derived the cumulative distribution function of Y. Since X is absolutely continuous, the function F_X is continuous as well, and hence we get the following expression for F_Y:

$$F_Y(y) = \begin{cases} F_X(2+\sqrt{y}) - F_X(2-\sqrt{y}), & \text{if } y > 0 \\ 0, & \text{otherwise} \end{cases}$$

This function is continuous (we only need to check it at $y = 0$), and differentiating everywhere except maybe at 0. By Theorem 2.25 the random variable Y is absolutely continuous, and we can obtain its probability density function by differentiating F_Y:

$$f_Y(y) = \begin{cases} \frac{1}{2\sqrt{y}} f_X(2+\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(2-\sqrt{y}), & \text{if } y > 0, \\ 0, & \text{otherwise.} \end{cases}$$

△

2.5. Further mathematical issues ♣

Absolutely continuous random variables.

The derivation if (2.26) for all Borel sets B involves the following reasoning. The defining property (2.25) of the density function f implies that (2.26) is true for sets B in the class $\mathcal{A} = \{(-\infty, b] : b \in \mathbb{R}\}$. This class $\mathcal{A}$ has two crucial properties: (i) $\mathcal{A}$ generates the Borel σ-algebra of $\mathbb{R}$, and (ii) $\mathcal{A}$ is closed under intersection. Verification of point (i) starts from the observation that any open interval (a, b) can be obtained from members of $\mathcal{A}$ with countably many set operations:

$$(a,b) = \Big(\bigcup_{n \ge 1} (-\infty, b - \tfrac{1}{n}\,] \Big) \setminus (-\infty, a].$$

Every open subset of $\mathbb{R}$ is a countable union of open intervals. Thus all open subsets are members of the σ-algebra generated by $\mathcal{A}$, and consequently so are all Borel sets.

Together, points (i) and (ii) are sufficiently strong to extend the validity of (2.26) from $B \in \mathcal{A}$ to all Borel sets B. The technical result that implies this is called the π-λ theorem (Theorems A.1.4 and A.1.5 in the appendix of [Dur10]). Same kind of reasoning justifies (2.40) in the multivariate case.

The precise condition on a cumulative distribution function F to have a density function is the following condition called *absolute continuity:* for every $\varepsilon > 0$ there exists $\delta > 0$ such that for any finite set of disjoint intervals (a_1, b_1), (a_2, b_2), $\ldots$, (a_N, b_N),

$$\sum_{i=1}^{N}(b_i - a_i) \le \delta \quad \text{implies} \quad \sum_{i=1}^{N}|F(b_i) - F(a_i)| \le \varepsilon. \tag{2.48}$$

(See Proposition 3.32 in [Fol99].) Absolute continuity can be difficult to check. The sufficient criterion of Theorem 2.25 is therefore convenient.

Theorem 2.25 is justified as follows. Let f be the derivative of F which is assumed to exist at all but at most countably many points of $\mathbb{R}$. Under this assumption, Theorem 6.3.10 of [Coh80] gives the relation

$$F(b) - F(a) = \int_a^b f(x)\,dx \tag{2.49}$$

for all real $a < b$, provided the integral on the right is finite. Note that the nonexistence of f at countably many points has no effect on the integral $\int_a^b f(x)\,dx$. We bound the integral by applying Fatou's lemma (Lemma 2.18 in [Fol99]) in the second step below:

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_a^b \lim_{n\to\infty} n\big(F(x+\tfrac{1}{n}) - F(x)\big)\,dx \\
&\le \varliminf_{n\to\infty} \int_a^b n\big(F(x+\tfrac{1}{n}) - F(x)\big)\,dx \\
&= \varliminf_{n\to\infty} \left(n\int_{a+\frac{1}{n}}^{b+\frac{1}{n}} F(x)\,dx \, - n\int_a^b F(x)\,dx \right) \\
&= \varliminf_{n\to\infty} \left(n\int_{b}^{b+\frac{1}{n}} F(x)\,dx \, - n\int_a^{a+\frac{1}{n}} F(x)\,dx \right) \\
&= F(b) - F(a) \le 1.
\end{aligned}$$

The last step above follows from the right-continuity of F. To complete the proof that f serves as the probability density function associated to F, we deduce relation (2.25) by letting $a \to -\infty$ in (2.49). The left-hand side of (2.49) converges to $F(b)$ by property (iii) of Theorem 2.20. The right-hand side of (2.49) converges to $\int_{-\infty}^b f(x)\,dx$ by the monotone convergence theorem (Theorem 2.14 in [Fol99]).

Exercises

Exercise 2.1. Roll a fair die repeatedly until you see the first even number. Let X be the number of rolls needed. Calculate the probabilities below.

(a) $P(X$ is even).

(b) $P(X$ is divisible by 2 or divisible by 3 or both).

(c) $P(X$ is divisible by 2 or divisible by 3 but not both).

Exercise 2.2. An urn contains 3 red balls and 2 yellow balls. You draw a ball without replacement until you get a red ball. Let X be the number of draws you made.

(a) Find the possible values and the probability mass function of X.

(b) Let μ_X be the probability distribution of X on $\mathbb{R}$. Imitating Example 2.10, give a formula for $\mu_X(B)$ for an arbitrary subset $B \subset \mathbb{R}$.

(c) Find the cumulative distribution function F of X.

Exercise 2.3. We have an urn which contains 5 marbles numbered $1, \dots, 5$. We sample two marbles without replacement.

(a) Identify the probability space $(\Omega, \mathcal{F}, P)$ corresponding to an ordered sample.

(b) Consider the collection $\mathcal{F}_1$ of events $A \in \mathcal{F}$ that have the following property: if $(a, b) \in A$ then $(b, a) \in A$. Show that this collection is also a σ-field.

(c) Denote by P_1 the restriction of the probability measure P to $\mathcal{F}_1$. Show that $(\Omega, \mathcal{F}_1, P_1)$ is a probability space. What experiment does it model?

(d) Find a function $X : \Omega \to \mathbb{R}$ that is not measurable with respect to $\mathcal{F}_1$.

Exercise 2.4. Parts (a)–(e) below describe five different probability spaces. In each case, define a function $X : \Omega \to \mathbb{R}$ such that X has probability distribution given by $P(X = 0) = \frac{1}{4}$ and $P(X = 1) = \frac{3}{4}$, or explain why this cannot be done.

(a) $\Omega = \{0, 1, 2, 3\}$, $P\{\omega\} = \frac{1}{4}$ for $\omega \in \Omega$.

(b) $\Omega = \{0, 1, 2, 3, 4\}$, $P\{\omega\} = \frac{1}{5}$ for $\omega \in \Omega$.

(c) $\Omega = \mathbb{Z}_{>0} = \{1, 2, 3, \dots\}$, $P\{\omega\} = 2^{-\omega}$ for $\omega \in \Omega$.

(d) $\Omega = [0, 1]$, $P(I) =$ length of I for intervals $I \subset \Omega$.

(e) $\Omega = \mathbb{R}$, $P((-\infty, 0]) = 0$, and $P((a, b)) = \int_a^b e^{-x}\,dx$ for intervals $(a, b) \subset (0, \infty)$.

Exercise 2.5. Roll a fair die twice and let X_1 and X_2 denote the outcomes of the rolls. Let $Y = X_1 \wedge X_2$ be the minimum of the two numbers. Find the probability mass function of Y. Check that your answer is a legitimate probability mass function.

Hint. The calculation is a little easier via $P(Y = k) = P(Y > k - 1) - P(Y > k)$.

Exercise 2.6. Let X be a random variable that satisfies $P(X = 1) = \frac{1}{2}$ and $P(a < X < b \,|\, X \neq 1) = \frac{1}{2}(b - a)$ for $0 \le a < b \le 2$. Find the cumulative distribution function of X. Present your answer as a case by case formula.

Exercise 2.7. Flip a fair coin. If it is heads, stop. If it is tails, roll a fair die once. Repeat. In other words, keep flipping the coin until the first heads, and after each coin flip that is tails, roll the die once. Find the probability that no sixes were rolled.

Hint. Let N be the number of flips needed for the first heads. Deduce the probabilities $P(\text{no sixes} \,|\, N = n)$ from the description of the game.

Exercise 2.8. A robot throws basketball hoops. Its accuracy improves after consecutive successful baskets. Fix a positive parameter α. Suppose the first throw succeeds with probability e^{-1}. If the first throw succeeded, the second throw succeeds with probability $e^{-2^{-\alpha}}$. And so on: after $k-1$ straight successes, the kth throw succeeds with probability $e^{-k^{-\alpha}}$. Let N be the number of throws until the first failure, including the failed throw. Calculate the probabilities $P(N=n)$ for $n \in \mathbb{Z}_{>0}$. For which values of α is there a positive probability that the robot never misses?

Exercise 2.9. Drop a uniformly random point on the triangle with vertices at $(0,0)$, $(5,0)$ and $(5,2)$. Let X be the x-coordinate of this random point. Find the cumulative distribution function of X.

Exercise 2.10. Let X be a random variable with cumulative distribution function F. Let $Y = X^2$. Find the cumulative distribution function F_Y of Y in terms of F.

Exercise 2.11. Let X be a random variable with cumulative distribution function F. The positive part of X is by definition $X^+(\omega) = X(\omega) \vee 0$ (the larger of $X(\omega)$ and zero). Find the cumulative distribution function F_{X^+} of X^+ in terms of F. Give an example of a cumulative distribution function F such that F is a continuous function but F_{X^+} is not.

Exercise 2.12. Let X be a random variable. Let s_n be a strictly decreasing sequence such that $s_n \to t$.

(a) Prove that the set convergence $\{X < s_n\} \searrow \{X \le t\}$ holds.

(b) Show by example that $\{X < s_n\} \searrow \{X < t\}$ can fail.

(c) Show by example that $\{X < s_n\} \searrow \{X < t\}$ can happen. **Hint.** While this may seem an odd task, it can in fact be satisfied by very trivial examples.

Exercise 2.13. Let F be the cumulative distribution function of a random variable X. Prove that $\lim_{t\to\infty} F(t) = 1$.

Exercise 2.14. Let X be a random variable with cumulative distribution function F.

(a) Show that the set $\{x \in \mathbb{R} : P(X = x) > 0\}$ is countable (that is, either finite or countably infinite).
Hint. How many points x can satisfy $P(X = x) \ge \frac{1}{n}$? For the proof you also need to know that a countable union of finite sets (and even of countable sets) is countable.

(b) How many discontinuities can F have?

Exercise 2.15. Prove Theorem 2.23.
Hint. Show that for $A = \cup_k \{s_k\}$ we have $P(X \in A) = 1$.

Exercise 2.16. Find a discrete random variable for which Theorem 2.23 cannot be applied.
Hint. You need to find a discrete random variable for which the set of possible values cannot be written as an increasing sequence.

Exercise 2.17. Let F and G be two cumulative distribution functions. Suppose that $G(x) = F(x)$ at each point $x \in \mathbb{R}$ at which F is continuous. Show that then $G(x) = F(x)$ at each $x \in \mathbb{R}$.

The point of the exercise is that the values $F(x)$ at continuity points are sufficient information for unique determination of the entire cumulative distribution function. This becomes important in Chapter 6.

Hint. From a correct solution to Exercise 2.14(b), one can conclude that continuity points of F are dense in $\mathbb{R}$.

Exercise 2.18. Let X be a random variable with cumulative distribution function F. Suppose that for all $t \in \mathbb{R}$, $F(t) = 0$ or $F(t) = 1$. Show that there exists a point $c \in \mathbb{R}$ such that $P(X = c) = 1$.

Hint. Work with $c = \inf\{t \in \mathbb{R} : F(t) = 1\}$.

Exercise 2.19. Let X be a random variable such that $P(X > 0) > 0$. Show that there exists a positive integer m such that $P(X \geq \frac{1}{m}) > 0$.

Exercise 2.20. Let X be a random variable. Show that $P(X < t)$ is a left-continuous function of the real variable t.

Exercise 2.21. Let X be uniformly distributed on the interval $[1, 5]$ and let Y be the distance from X to the nearest endpoint of the interval. Find the cumulative distribution function and probability density function of Y. Identify the distribution of Y by name.

Exercise 2.22. Suppose that when Rick has a car accident, the cost of the repair to this car in dollars is uniformly distributed on the interval $[100, 1300]$. He has an insurance policy with an 800 dollar deductible. This means that the insurance company pays for any damages that go over 800 dollars. Let Y be the cost of the repair to the car, and let X be the amount that Rick pays. Express X as a function of Y. Find the cumulative distribution functions of both X and Y. Find $P(X = x)$ for all points x for which the probability is strictly positive.

Exercise 2.23. Suppose we know the following about a random variable X:

$$\begin{aligned} &P(X < 1) = P(X > 2) = 0, \\ &P(X = 1) = P(X = 2), \\ \text{and} \quad &P(a < X < b) = \frac{b^3 - a^3}{14} \quad \text{for } 1 \leq a < b \leq 2. \end{aligned}$$

Find the cumulative distribution function F of X.

Exercise 2.24. In the following functions, λ and θ are fixed positive real parameters, and c is an unknown constant. In each case, identify those values of c for which the function f_i is a probability density function on $\mathbb{R}$. In some cases no value of c will do. In some cases some values of λ or θ may also have to be discarded. Explain your reasoning.

(a) $f_1(x) = \begin{cases} 0, & x \leq 0 \\ ce^{-\lambda x}, & x > 0. \end{cases}$

(b) $f_2(x) = \begin{cases} 0, & x \leq 0 \\ \dfrac{c}{\Gamma(\theta)} x^{\theta - 1} e^{-\lambda x}, & x > 0. \end{cases}$

Note: the *gamma function* is by definition $\Gamma(t) = \displaystyle\int_0^\infty x^{t-1} e^{-x}\, dx$ for $t > 0$.

(c) $f_3(x) = \begin{cases} 0, & x \le 1 \\ cx^{-\theta}, & x > 1. \end{cases}$

(d) $f_4(x) = \begin{cases} 0, & x < 0 \text{ or } x > 2\pi \\ c\sin x & x \in [0, 2\pi]. \end{cases}$

(e) $f_5(x) = \begin{cases} 0, & x < 0 \text{ or } x > \pi \\ c\sin x & x \in [0, \pi]. \end{cases}$

(f) $f_6(x) = \dfrac{c}{1+x^2}$ for all $x \in \mathbb{R}$.

Some of these density functions have names: f_1 is density of the *exponential distribution*, f_2 of the *gamma distribution*, and f_6 of the *Cauchy distribution*.

Exercise 2.25. Let (X, Y) be a uniformly random point on the 2×2 square with four corners at $(0, 0)$, $(2, 0)$, $(2, 2)$ and $(0, 2)$. Let R be the distance of (X, Y) to the nearest corner. Find the cumulative distribution function and the probability density function of R.

Hint. Number the corners 1 through 4. Let $r \ge 0$. Center a disk of radius r at each corner. For corner i, let B_i be the portion of this disk that lies inside the square $\Omega = [0, 2] \times [0, 2]$. Then $P(R \le r) = P(B_1 \cup B_2 \cup B_3 \cup B_4)$. See Figure 1.

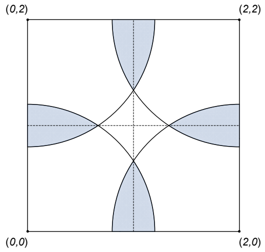

Figure 1. Illustration for Exercise 2.25. The overlaps of the disks inside the square are shaded.

Exercise 2.26. Let X be a random variable with density function f. Assume that f is continuous at $x = a$. Prove the limit (2.35) by making the argument described in Remark 2.29 rigorous.

Exercise 2.27. Let $F(x, y)$ be the joint cumulative distribution function of the random vector (X, Y). Show that if $a \le b$, $c \le d$ then

(1) $F(a, c) \le F(b, d)$

(2) $F(a, d) + F(b, c) \le F(a, c) + F(b, d)$.

Exercise 2.28. Let $F(x, y)$ be the joint cumulative distribution function of the random vector (X, Y). Evaluate the following limits:

(a) $\lim_{x,y\to\infty} F(x, y)$

(b) $\lim_{x\to\infty} F(x, y)$ (Here $y \in \mathbb{R}$ is fixed.)

(c) $\lim_{x\to-\infty} F(x,y)$ (Here $y \in \mathbb{R}$ is fixed.)

Exercise 2.29. Roll a fair die twice and let X_1 and X_2 denote the outcomes of the rolls. Let $Y = X_1 \wedge X_2$ be the minimum of the two numbers. Find the joint probability mass function $p_{X_1,Y}$ of (X_1, Y). Derive the marginal probability mass functions p_{X_1} and p_Y from $p_{X_1,Y}$.

Exercise 2.30. Let c be a positive constant and

$$f(x,y) = \begin{cases} cxe^{-2x(1+y)}, & \text{if } x > 0 \text{ and } y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Which value of c makes f a probability density function on $\mathbb{R}^2$?

(b) Let (X, Y) have joint probability density function f given above, with the right constant c from part (a). Find the marginal density functions f_X of X and f_Y of Y.

Exercise 2.31. Let X and Y be two random variables with cumulative distribution functions F_X and F_Y, respectively.

(a) Prove or disprove: if $F_X(s) = F_Y(s)$ for all rationals s then $X \overset{d}{=} Y$.

(b) Prove or disprove: if $F_X(k) = F_Y(k)$ for all integers k then $X \overset{d}{=} Y$.

Exercise 2.32. Suppose that X is a random variable with cumulative distribution function $F_X(x) = 1 - e^{-x}$ for $x \geq 0$ and $F_X(x) = 0$ for $x < 0$. Let $Y = \log X$.

(a) Find the cumulative distribution function of Y.

(b) Show that Y is absolutely continuous and find its probability density function.

Exercise 2.33. Let X be a uniform random variable on $[-1, 2]$. Show that $Y = |X - 1|$ is absolutely continuous and find its density.

Exercise 2.34. Suppose that X is uniformly distributed on $[-2\pi, 2\pi]$. Find the probability density function of $Y = \tan X$.

Exercise 2.35. Recall that $\lfloor x \rfloor$ gives the largest integer not bigger than x. The function $\{x\}$ is defined as $\{x\} = x - \lfloor x \rfloor$ and it is called the fractional part of x. (E.g. $\{3.6\} = 0.6$, $\{6\} = 0$, $\{-1.2\} = 0.8$.) Suppose that X is uniformly distributed on $[0, 1]$.

(a) Find the probability density function of $Y = \left\{\frac{1}{X}\right\}$.

(b) Show that $Z = \lfloor X \rfloor$ is discrete and find its probability mass function.

Exercise 2.36. Let X be an absolutely continuous random variable with density function f_X. Suppose that g is strictly increasing and continuously differentiable. Derive the probability density function of the random variable $Y = g(X)$.
Hint. Find the cumulative distribution function of Y first and then differentiate. You will need the derivative of the inverse function.

Exercise 2.37. Suppose the cumulative distribution function F_X of the random variable X is an invertible function. Let U be a uniform random variable on $[0, 1]$. Show that $Y = F^{-1}(U)$ has the same distribution as X.

Exercise 2.38. Suppose that X is a discrete random variable with probability mass function p_X and U is a uniform random variable on $[0, 1]$. Find a function g so that $g(U)$ has the same distribution as X.

Exercise 2.39.* Suppose that X is a random variable with cumulative distribution function F and U is a uniform random variable on $[0, 1]$. Define the function $g : (0, 1) \to \mathbb{R}$ as

$$g(x) = \sup\{y : F(y) < x\}.$$

Prove that $g(U)$ has the same distribution as X.

The result of this problem shows that any one dimensional distribution can be generated as an appropriate function of a uniform random variable.

Chapter 3

Independent and dependent random variables

3.1. Independent random variables

We have already seen the definition of independent random variables in Definition 1.55. We restate an equivalent formulation, and then state a version for random vectors.

Definition 3.1. Let $X_1, X_2, \dots, X_n$ be random variables defined on the same probability space. Then $X_1, X_2, \dots, X_n$ are independent if

$$P(X_1 \in B_1, X_2 \in B_2, \dots, X_n \in B_n) = \prod_{i=1}^{n} P(X_i \in B_i) \tag{3.1}$$

for all Borel subsets $B_1, B_2, \dots, B_n$ of the real line.

This definition looks more restrictive than Definition 1.55, since there we only require (3.1) for sets of the form $B_k = (-\infty, c_k]$. However, it can be shown that if (3.1) holds for these sets, then it also holds for all Borel sets. (Proving this statement requires tools from measure theory that are beyond the scope of these notes.)

Definition 3.2. Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be random vectors defined on the same probability space. Let $\mathbf{X}_i$ be $\mathbb{R}^{d_i}$-valued. Then $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent if

$$P(\mathbf{X}_1 \in B_1, \mathbf{X}_2 \in B_2, \dots, \mathbf{X}_n \in B_n) = \prod_{k=1}^{n} P(\mathbf{X}_k \in B_k) \tag{3.2}$$

for all Borel subsets $B_i \subset \mathbb{R}^{d_i}$, $i = 1, \dots, n$.

Definition 3.2 actually covers Definition 3.1 also since a one-dimensional random vector is a random variable. Hence there is no point in always making separate statements for random variables and random vectors. In the sequel we sometimes give only the random vector statement with the understanding that random variables are also covered as a special case.

Note that both definitions only use the joint distributions of the random variables (or vectors) in question.

As for the independence of events, the independence of infinitely many random variables reduces back to the finite definition.

Definition 3.3. Let $\{\mathbf{X}_k\}_{k\in\mathbb{Z}_{>0}}$ be an infinite sequence of random vectors defined on some probability space $(\Omega, \mathcal{F}, P)$. Then the random vectors $\{\mathbf{X}_k\}_{k\in\mathbb{Z}_{>0}}$ are independent if, for each $n \in \mathbb{Z}_{>0}$, the random vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent. △

The general definitions above are usually not practical criteria for checking independence. In concrete situations we turn to the joint cumulative distribution function, the probability mass function and the probability density function. For each of these, a product property is equivalent to the independence of the random vectors and variables. The criterion in terms of the cumulative distribution function stated below is valid for all random variables and vectors.

Theorem 3.4. *Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be random vectors defined on the same probability space. For $1 \le i \le n$ let d_i be the dimension of $\mathbf{X}_i$ and set $d = d_1 + \dots + d_n$. Define the d-dimensional random vector $\mathbf{Y} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ by putting all the coordinates of the $\mathbf{X}_i$s together. Denote the joint cumulative distribution functions of these random vectors by $F_{\mathbf{Y}}$, $F_{\mathbf{X}_1}$, …, $F_{\mathbf{X}_n}$. Then $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent if and only if*

$$F_{\mathbf{Y}}(\mathbf{x}_1, \dots, \mathbf{x}_n) = F_{\mathbf{X}_1}(\mathbf{x}_1) \cdot F_{\mathbf{X}_2}(\mathbf{x}_2) \cdots F_{\mathbf{X}_n}(\mathbf{x}_n) \tag{3.3}$$

for all vectors $\mathbf{x}_1 \in \mathbb{R}^{d_1}, \mathbf{x}_2 \in \mathbb{R}^{d_2}, \dots, \mathbf{x}_n \in \mathbb{R}^{d_n}$.

Exercise 3.2 asks you to prove half of this theorem in a special case.

Next we treat discrete and jointly absolutely continuous random variables separately.

Independence of discrete random variables and vectors.

Theorem 3.5. *Discrete random vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent if and only if*

$$P(\mathbf{X}_1 = \mathbf{k}_1, \mathbf{X}_2 = \mathbf{k}_2, \dots, \mathbf{X}_n = \mathbf{k}_n) = \prod_{i=1}^{n} P(\mathbf{X}_i = \mathbf{k}_i) \tag{3.4}$$

for all choices $\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n$ of possible values of the random vectors.

If (3.4) is assumed for possible values of the random vectors, it then actually holds for all choices of $\mathbf{k}_1, \dots, \mathbf{k}_n$ because if any of the values is not possible for one of the random vectors, both sides of (3.4) equal zero.

Proof. The task is to show that for discrete random vectors Definition 3.2 is equivalent to the condition in Theorem 3.5. If we assume Definition 3.2, then identity (3.4) is the special case of (3.2) obtained by taking $B_i = \{\mathbf{k}_i\}$ for $i = 1, \dots, n$.

Conversely, assume that (3.4) holds for all choices of $\mathbf{k}_1, \dots, \mathbf{k}_n$. Let B_i be arbitrary subsets of $\mathbb{R}^{d_i}$ for $i = 1, \dots, n$ where, as in Definition 3.2, d_i is the dimension of the vector $\mathbf{X}_i$. The proof goes by decomposing the probability. In the sums below the indices $\mathbf{k}_1, \dots, \mathbf{k}_n$ range over those possible values of the random vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$ that satisfy the constraints $\mathbf{k}_1 \in B_1, \dots, \mathbf{k}_n \in B_n$.

$$\begin{aligned}
&P\{\mathbf{X}_1 \in B_1, \mathbf{X}_2 \in B_2, \dots, \mathbf{X}_n \in B_n\} \\
&= \sum_{\mathbf{k}_1 \in B_1,\, \mathbf{k}_2 \in B_2, \dots,\, \mathbf{k}_n \in B_n} P\{\mathbf{X}_1 = \mathbf{k}_1, \mathbf{X}_2 = \mathbf{k}_2, \dots, \mathbf{X}_n = \mathbf{k}_n\} \\
&= \sum_{\mathbf{k}_1 \in B_1,\, \mathbf{k}_2 \in B_2, \dots,\, \mathbf{k}_n \in B_n} P\{\mathbf{X}_1 = \mathbf{k}_1\} P\{\mathbf{X}_2 = \mathbf{k}_2\} \cdots P\{\mathbf{X}_n = \mathbf{k}_n\} \\
&= \Bigg(\sum_{\mathbf{k}_1 \in B_1} P\{\mathbf{X}_1 = \mathbf{k}_1\} \Bigg) \Bigg(\sum_{\mathbf{k}_2 \in B_2} P\{\mathbf{X}_2 = \mathbf{k}_2\} \Bigg) \cdots \Bigg(\sum_{\mathbf{k}_n \in B_n} P\{\mathbf{X}_n = \mathbf{k}_n\} \Bigg) \\
&= P\{\mathbf{X}_1 \in B_1\} \cdot P\{\mathbf{X}_2 \in B_2\} \cdots P\{\mathbf{X}_n \in B_n\}.
\end{aligned}$$

The assumed condition (3.4) justified the second equality. Thereby we verified (3.2) and independence of the random vectors $\mathbf{X}_i$. □

We can strengthen the first direction of Theorem 3.5 as follows.

Theorem 3.6. *Suppose that for the discrete random vectors $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ there exists non-negative functions $g_j, 1 \le j \le n$ so that*

$$P(\mathbf{X}_1 = \mathbf{k}_1, \mathbf{X}_2 = \mathbf{k}_2, \dots, \mathbf{X}_n = \mathbf{k}_n) = \prod_{i=1}^{n} g_i(\mathbf{k}_i) \tag{3.5}$$

for all choices $\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_n$. Then $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ are independent, and there exist positive constants $c_j, 1 \le j \le n$ so that $\prod_{j=1}^{n} c_j = 1$, and $g_j(\mathbf{k}) = c_j P(\mathbf{X}_1 = \mathbf{k})$ for all $\mathbf{k}$, $1 \le j \le n$. △

Exercise 3.5 gives the proof of this theorem in a simpler setup. We show the proof in the case of two random variables.

Proof of Theorem 3.6 for two random variables. We assume that there are non-negative functions g_1, g_2 so that for the discrete random variables X_1, X_2 we have

$$P(X_1 = k_1, X_2 = k_2) = g_1(k_1) g_2(k_2) \tag{3.6}$$

for all $k_1, k_2 \in \mathbb{R}$. Let S be the union of the set of possible values of X_1 and X_2. This is a countable set, and $P(X_1 = k_1, X_2 = k_2) = 0$ if k_1 or k_1 is not in S. We sum (3.6) for all $k_1, k_2 \in S$:

$$\sum_{k_1, k_2 \in S} P(X_1 = k_1, X_2 = k_2) = \sum_{k_1, k_2 \in S} g_1(k_1) g_2(k_2) = \Bigg(\sum_{k_1 \in S} g_1(k_1) \Bigg) \Bigg(\sum_{k_2 \in S} g_2(k_2) \Bigg).$$

The second step can be checked by expanding the terms in the last product. (This identity works even for infinitely many terms, if all terms are non-negative.) Since S contains all possible values of X_1 and X_2, we have

$$\sum_{k_1, k_2 \in S} P(X_1 = k_1, X_2 = k_2) = 1.$$

Let

$$c_1 = \sum_{k_1 \in S} g_1(k_1), \qquad c_2 = \sum_{k_1 \in S} g_2(k_2).$$

These are sums of non-negative numbers, and their product is equal to 1, so c_1 and c_2 are both positive real numbers. Now fix k_1 and sum (3.6) for all $k_2 \in S$:

$$\begin{aligned} P(X_1 = k_1) &= \sum_{k_2 \in S} P(X_1 = k_1, X_2 = k_2) = \sum_{k_2 \in S} g_1(k_1) g_2(k_2) \\ &= g_1(k_1) \left(\sum_{k_2 \in S} g_2(k_2) \right) = c_2 g_1(k_1). \end{aligned}$$

Since $c_2 = \frac{1}{c_1}$ we get that

$$P(X_1 = k_1) = \frac{1}{c_1} g_1(k_1).$$

We can similarly prove that $P(X_2 = k_2) = \frac{1}{c_2} g_2(k_2)$ which proves the second part of the statement. Using $c_1 c_2 = 1$ again we also get that

$$P(X_1 = k_1, X_2 = k_2) = \frac{1}{c_1} g_1(k_1) \frac{1}{c_2} g_2(k_2) = P(X_1 = k_1) P(X_2 = k_2)$$

for all $k_1, k_2 \in S$, which shows that X_1, X_2 are independent. □

Example 3.7. We flip a coin n times. Let X_k for $1 \le k \le n$ be the indicator of the event that the kth flip is tails. Then $X_1, \dots, X_n$ are independent.

We use the probability space introduced in Example 1.13: $\Omega = \{0,1\}^n$, $P(\omega) = 2^{-n}$ for $\omega \in \Omega$. Note that $X_k(\omega) = s_k$ for $\omega = (s_1, \dots, s_n)$. The possible values for X_k are 0 and 1, so we need to check that

$$P(X_1 = k_1, \dots, X_n = k_n) = \prod_{j=1}^{n} P(X_j = k_j) \tag{3.7}$$

holds for all choices of $(k_1, \dots, k_n) \in \{0,1\}^n$. We have

$$P(X_1 = k_1, \dots, X_n = k_n) = P(\omega = (k_1, \dots, k_n)) = 2^{-n},$$

since the values of $X_1, \dots X_n$ identify the outcomes of the n coin flips. In Example 1.51 we checked that $P(X_j = 1) = \frac{1}{2}$, which also gives $P(X_j = 0) = \frac{1}{2}$. But this shows that the right side of (3.7) is equal to $(\frac{1}{2})^n = 2^{-n}$, which proves (3.7) for all choices of $(k_1, \dots, k_n) \in \{0,1\}^n$. This means that $X_1, \dots, X_n$ are indeed independent. △

Example 3.8. (Sampling with and without replacement.) There is an urn with m balls labeled 1 through m. You reach in, draw a ball, and record the number you drew. This step is repeated k times. There are two ways to carry out this repeated sampling.

- In *sampling with replacement* the ball is put back into the urn after each draw. The contents of the urn are the same for each draw.
- In *sampling without replacement* each drawn ball is put aside, in other words, not back into the urn. So each ball can be drawn at most once. In this setting necessarily $k \le m$.

The fundamental assumption is that each sample of k numbers is equally likely.

Let $X_1, X_2, \dots, X_k$ denote the outcomes of the successive draws. We derive their joint distribution, marginal distributions, and investigate their independence. Intuitively the independence issue should be clear: in sampling with replacement there is no influence from one draw to the next, while in sampling without replacement each draw constrains the options available for later draws.

Sampling with replacement. The sample space is the Cartesian product space of ordered k-tuples from the set $\{1, \dots, m\}$:

$$\Omega = \{1, \dots, m\}^k = \{\omega = (s_1, \dots, s_k) : \text{each } s_i \in \{1, \dots, m\}\}$$

with $\#\Omega = m^k$. The random variables are the coordinates: $X_j(\omega) = s_j$.

Let us check that $X_1, X_2, \dots, X_k$ are independent. First compute, for any $x \in \{1, \dots, m\}$,

$$P(X_j = x) = \frac{m^{k-1}}{m^k} = \frac{1}{m}$$

by counting favorable arrangements (the jth draw gives x, the other $k-1$ are unrestricted). Let $x_1, x_2, \dots, x_k \in \{1, \dots, m\}$. Note that specifying the values of all $X_1, X_2, \dots, X_k$ amounts to specifying a unique outcome from Ω.

$$P\big(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k\big) = \frac{1}{\#\Omega} = \frac{1}{m^k} = \left(\frac{1}{m}\right)^k = \prod_{j=1}^{k} P(X_j = x_j).$$

We have verified the criterion in Theorem 3.5 and thereby checked that $X_1, \dots, X_k$ are mutually independent.

Sampling without replacement. Now $k \le m$ and the sample space Ω is the space of k-tuples of distinct entries from $\{1, \dots, m\}$:

$$\Omega = \{\omega = (s_1, \dots, s_k) : \text{each } s_i \in \{1, \dots, m\} \text{ and all } s_i \text{ are distinct}\}$$

with $\#\Omega = m(m-1)\cdots(m-k+1)$. The random variables are again the coordinates $X_j(\omega) = s_j$. As above, specifying the values of all $X_1, X_2, \dots, X_k$ amounts to specifying a unique outcome from Ω, and so

$$P\big(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k\big) = \frac{1}{\#\Omega} = \frac{1}{m(m-1)\cdots(m-k+1)},$$

provided $x_1, x_2, \dots, x_k$ are distinct elements of the set $\{1, \dots, m\}$.

We derive the probability mass functions of the individual draws. For the first draw, counting favorable arrangements,

$$P(X_1 = x) = \frac{1 \cdot (m-1)(m-2)\cdots(m-k+1)}{m(m-1)\cdots(m-k+1)} = \frac{1}{m}.$$

For the second draw

$$P(X_2 = x) = \frac{(m-1) \cdot 1 \cdot (m-2)(m-3)\cdots(m-k+1)}{m(m-1)\cdots(m-k+1)} = \frac{1}{m}$$

because the first draw can be anything other than x ($m-1$ choices), the second draw is restricted to x (1 choice), and the third draw has $m-2$ alternatives, and so on. The general pattern is

$$P(X_j = x) = \frac{(m-1)\cdots(m-j+1)\cdot 1\cdot(m-j)\cdots(m-k+1)}{m(m-1)\cdots(m-k+1)} = \frac{1}{m}.$$

We can see already that independence fails in sampling without replacement, since the joint probability mass function of $X_1, \dots, X_k$ is not the product of the marginal probability mass functions. In fact, these random variables are not even pairwise independent since for $i \neq j$, $P(X_i = 1, X_j = 1) = 0$ but $P(X_i = 1)P(X_j = 1) = m^{-2} \neq 0$.

Marginal distributions do not distinguish between sampling with and without replacement. The marginal distribution of the jth draw is the same for both, namely $P(X_j = x) = \frac{1}{m}$ for each $x \in \{1, \dots, m\}$. △

Definition 3.9. Random variables $X_1, X_2, X_3, \dots$ are **independent and identically distributed** (abbreviated **i.i.d.**) if they are independent and each X_k has the same probability distribution. That is, $X_k \overset{d}{=} X_\ell$ for any two indices k, ℓ. △

A sequence of random variables can also be called a *stochastic process*. In case the random variables are i.i.d., the process is called an *i.i.d. process*.

Example 3.10. In Example 3.8 we found that in sampling with replacement, the draws are i.i.d. In sampling without replacement, the draws have identical distributions but they are not independent. △

Independence of jointly absolutely continuous random variables.

Theorem 3.11. *Let $X_1, \dots, X_d$ be random variables on the same sample space. Assume that for each $j = 1, 2, \dots, d$, X_j has density function f_{X_j}.*

(a) *If $X_1, \dots, X_d$ have joint density function f given by*

$$f(x_1, x_2, \dots, x_d) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_d}(x_d) \tag{3.8}$$

then $X_1, \dots, X_d$ are independent.

(b) *Suppose $X_1, \dots, X_d$ are independent. Then they are jointly absolutely continuous with joint density function*

$$f(x_1, x_2, \dots, x_d) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_d}(x_d).$$

Proof. Let $B_1, \dots, B_d$ be Borel subsets of $\mathbb{R}$. Part (a).

$$\begin{aligned} P(X_1 \in B_1, \dots, X_d \in B_d) &= \int\limits_{B_1 \times \cdots \times B_d} f(\mathbf{x})\, d\mathbf{x} \\ &= \int\limits_{B_1}\int\limits_{B_2} \cdots \int\limits_{B_d} f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_d}(x_d)\, dx_1\, dx_2 \cdots dx_d \\ &= \int\limits_{B_1} f_{X_1}(x_1)\, dx_1 \int\limits_{B_2} f_{X_2}(x_2)\, dx_2 \cdots \int\limits_{B_d} f_{X_d}(x_d)\, dx_d \\ &= P(X_1 \in B_1)\, P(X_2 \in B_2) \cdots P(X_d \in B_d). \end{aligned}$$

Part (b). Now independence is assumed.

$$\begin{aligned}
&P(X_1 \in B_1, \dots, X_d \in B_d) \\
&\qquad = P(X_1 \in B_1)\, P(X_2 \in B_2) \cdots P(X_d \in B_d) \\
&\qquad = \int\limits_{B_1} f_{X_1}(x_1)\, dx_1 \int\limits_{B_2} f_{X_2}(x_2)\, dx_2 \;\cdots \int\limits_{B_d} f_{X_d}(x_d)\, dx_d \\
&\qquad = \int\limits_{B_1 \times B_2 \times \cdots \times B_d} f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_d}(x_d)\, dx_1\, dx_2 \cdots dx_d.
\end{aligned}$$

This calculation shows that, with $\mathbf{X} = (X_1, \dots, X_d)$ and $f(\mathbf{x}) = \prod_{i=1}^d f_i(x_i)$, the identity

$$P(\mathbf{X} \in B) = \int_B f(\mathbf{x})\, d\mathbf{x}$$

holds for sets B of the Cartesian product type $B = B_1 \times \cdots \times B_d$. This is sufficiently strong to imply the identity for all Borel sets $B \subset \mathbb{R}^d$. □

In Example 3.12 below we build independence into the model, while in the subsequent Example 3.13 we discover the independence of random variables.

Example 3.12. (Independent exponential random variables.) Suppose that X and Y are independent absolutely continuous random variables with densities

$$f_X(t) = \begin{cases} \lambda e^{-\lambda t}, & t \ge 0, \\ 0, & t < 0, \end{cases} \qquad f_Y(t) = \begin{cases} \mu e^{-\mu t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

(These are called exponential distributions with parameter λ and μ, respectively.) Find joint density $f_{X,Y}$, calculate the probability $P(X < Y)$, and derive the distribution of the minimum $T = \min(X, Y)$.

Independence of X and Y then implies that

$$f_{X,Y}(s,t) = f_X(s) f_Y(t) = \begin{cases} \lambda\mu e^{-\lambda s - \mu t}, & s, t \ge 0 \\ 0, & s < 0 \text{ or } t < 0. \end{cases}$$

We can compute $P(X < Y)$ by integrating the joint density on the set $\{(x,y) : x < y\}$:

$$\begin{aligned}
P(X < Y) &= \iint_{s<t} f_{X,Y}(s,t)\, ds\, dt = \int_0^\infty \lambda e^{-\lambda s} \Bigl(\int_s^\infty \mu e^{-\mu t}\, dt \Bigr)\, ds \\
&= \int_0^\infty \lambda e^{-\lambda s} \cdot e^{-\mu s}\, ds = \frac{\lambda}{\lambda + \mu} \int_0^\infty (\lambda + \mu) e^{-(\lambda+\mu)s}\, ds \qquad (3.9) \\
&= \frac{\lambda}{\lambda + \mu}.
\end{aligned}$$

Because $T = \min(X, Y)$ is defined as a minimum, it is convenient to compute its tail probability $P(T > t)$ and use independence. For $t > 0$:

$$\begin{aligned}
P(T > t) &= P(\min(X,Y) > t) = P(X > t,\, Y > t) = P(X > t) P(Y > t) \\
&= e^{-\lambda t} \cdot e^{-\mu t} = e^{-(\lambda+\mu)t}.
\end{aligned}$$

This shows that $P(T \le t) = 1 - e^{-(\lambda+\mu)t}$ for $t \ge 0$, and since $T \ge 0$ we also have $P(T \le t) = 0$ for $t < 0$. Differentiating this we get that the probability density function of T is

$$f_T(t) = \begin{cases} (\lambda+\mu)e^{-(\lambda+\mu)t}, & t \ge 0, \\ 0, & t < 0. \end{cases}$$

Note that the probability density function of T has the same form as the probability density function of X (and Y), with $\lambda + \mu$ in place of λ (and μ). △

Example 3.13. (Polar coordinates.) Let (X, Y) be a uniform random point on a disk Ω centered at $(0, 0)$ with radius r_0. Let (R, Θ) denote the polar coordinates of the point (X, Y). We find the joint and marginal distributions of (R, Θ) and show that R and Θ are independent.

We first compute the joint cumulative distribution function $F_{R,\Theta}$. By definition $0 \le R \le r_0$ and $\Theta \in [0, 2\pi)$. Thus we need to compute $F_{R,\Theta}(u, v) = P(R \le u, \Theta \le v)$ for $u \in [0, r_0]$ and $v \in [0, 2\pi)$. Let $(r(x, y), \theta(x, y))$ denote the functions that give the polar coordinates (r, θ) of a point (x, y). Then the set

$$A_{u,v} = \{(x, y) : r(x, y) \le u, \theta(x, y) \le v\}$$

is a circular sector of radius u, bounded by the angles 0 and v, and has area $\frac{1}{2}u^2 v$. Thus the probability that the random point (X, Y) lies in $A_{u,v}$ is

$$F_{R,\Theta}(u, v) = P(R \le u, \Theta \le v) = P((X, Y) \in A_{u,v}) = \frac{\frac{1}{2}u^2 v}{r_0^2 \pi} = \frac{u^2 v}{2r_0^2 \pi}.$$

By taking u to the upper bound r_0 of R gives us the cumulative distribution function of Θ: for $v \in [0, 2\pi)$,

$$F_\Theta(v) = P(\Theta \le v) = P(R \le r_0, \Theta \le v) = \frac{v}{2\pi}. \tag{3.10}$$

In Example 2.19 we computed the cumulative distribution function of R: $F_R(u) = u^2/r_0^2$ for $u \in [0, r_0)$. From these formulas one can check that $F_{R,\Theta}(u, v) = F_R(u)F_\Theta(v)$ holds for all $(u, v) \in \mathbb{R}^2$. Thus by Theorem 3.4, R and Θ are independent.

We can also see the independence from the density functions. By (2.41), the joint density function in the range $0 < r < r_0$ and $0 < \theta < 2\pi$ comes from

$$f_{R,\Theta}(r, \theta) = \frac{\partial^2}{\partial r \partial \theta} F_{R,\Theta}(r, \theta) = \frac{\partial^2}{\partial r \partial \theta}\left(\frac{r^2 \theta}{2r_0^2 \pi}\right) = \frac{r}{r_0^2 \pi}. \tag{3.11}$$

Outside of the range $(r, \theta) \in (0, r_0) \times (0, 2\pi)$ we can set $f_{R,\Theta} = 0$.

Differentiation of (3.10) gives

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 < \theta < 2\pi \\ 0, & \text{otherwise.} \end{cases}$$

From both F_Θ and f_Θ we see that the angle Θ is uniformly distributed on $[0, 2\pi)$. The density function of R was derived in Example 2.26:

$$f_R(r) = \begin{cases} \frac{2r}{r_0^2}, & 0 < r < r_0 \\ 0, & \text{otherwise.} \end{cases}$$

The formulas confirm that $f_{R,\Theta}(r,\theta) = f_R(r)f_\Theta(\theta)$ for all $(r,\theta) \in \mathbb{R}^2$, and threreby verify again the independence of R and Θ. △

Finding independence.

For the next proof we introduce the notion of the *inverse image of a set*. If $f : S \to T$ is a function from a space S into a space T, then for any subset $B \subset T$ we define its inverse image $f^{-1}(B)$ by

$$f^{-1}(B) = \{x \in S : f(x) \in B\}.$$

In plain words, $f^{-1}(B)$ is the set of those points in S that f maps into the set B. We have already used this notion a number of times. For a random variable X and a set B, the event $\{X \in B\}$ is precisely the inverse image $X^{-1}(B)$.

Theorem 3.14. *Suppose $X_1, \dots, X_n$ are independent random variables and for each index i, g_i is a function on the range of X_i. Then the random variables $g_1(X_1)$, $g_2(X_2), \dots, g_n(X_n)$ are independent.*

One needs to assume that $g_1, \dots, g_n$ are measurable, but this does not come up usually in applications.

Proof. This is now immediate from the definitions.

$$\begin{aligned}
&P\{g_1(X_1) \in B_1,\, g_2(X_2) \in B_2, \dots,\, g_n(X_n) \in B_n\} \\
&= P\{X_1 \in g_1^{-1}(B_1),\, X_2 \in g_2^{-1}(B_2), \dots, X_n \in g_n^{-1}(B_n)\} \\
&= \prod_{i=1}^n P\{X_i \in g_i^{-1}(B_i)\} = \prod_{i=1}^n P\{g_i(X_i) \in B_i\}.
\end{aligned}$$ □

With this theorem we can extend the independence of sampling with replacement to some situations where outcomes are not equally likely.

Example 3.15. An urn contains 5 red, 3 yellow and 2 green balls. A ball is sampled n times with replacement. Let $Y_k \in \{\texttt{r}, \texttt{y}, \texttt{g}\}$ be the color of the kth draw. Show that the random variables $\{Y_k\}_{k=1}^n$ are i.i.d.

Number the balls from 1 to 10. The assumption we make is that among the draws of balls all outcomes are equally likely. Let X_k be the number of the kth draw. By Examples 3.8 and 3.10, the random variables $\{X_k\}_{k=1}^n$ are i.i.d.

Next, paint balls 1-5 red, balls 6-8 yellow, and balls 9-10 green. Mathematically speaking, define the function

$$g(i) = \begin{cases} \texttt{r}, & i \in \{1, \dots, 5\} \\ \texttt{y}, & i \in \{6, 7, 8\} \\ \texttt{g}, & i \in \{9, 10\} \end{cases}$$

from the numbers on the balls to the colors. Then $Y_k = g(X_k)$ is the color of the kth draw. By Theorem 3.14, $\{Y_k\}_{k=1}^n$ are independent. Each Y_k has the same probability mass function

$$p_Y(\texttt{r}) = \tfrac{5}{10}, \quad p_Y(\texttt{y}) = \tfrac{3}{10}, \quad \text{and} \quad p_Y(\texttt{g}) = \tfrac{2}{10}$$

and so they are equal in distribution. In summary, $\{Y_k\}_{k=1}^n$ are i.i.d. △

As for events in Theorem 1.53, it is also true for random variables that if new random variables are formed from old independent ones so that separate collections are used, then the new random variables inherit the independence. We state the theorem that imitates the formulation of Theorem 1.53.

Theorem 3.16. *Let $\{X_k\}_{k\geq 1}$ be a finite or infinite sequence of independent random variables. Let $0 = k_0 < k_1 < \cdots < k_n$ be integers. Let $g_1, \dots, g_n$ be functions such that g_j is defined on the range of the random vector $(X_{k_{j-1}+1}, \dots, X_{k_j})$. Define new random variables $Y_j = g_j(X_{k_{j-1}+1}, \dots, X_{k_j})$ for $j = 1, \dots, n$. Then the random variables $Y_1, \dots, Y_n$ are independent.*

Example 3.17. Let $\{X_i\}$ be the outcomes of repeated independent fair coin flips (with 1 for tails, 0 for heads). For integers $m < n$ define $S_{m,n} = \sum_{i=m+1}^{n} X_i$, the number of tails in the coin flips $m+1, m+2, \dots, n$. Then for any integers $0 = n_0 < n_1 < \cdots < n_k$, the random variables S_{n_0,n_1}, $S_{n_1,n_2}, \dots, S_{n_{k-1},n_k}$ are independent. △

By adapting the examples above, we can represent any i.i.d. random variables with finitely many possible values which all have rational probabilities by equally likely draws with replacement (Exercise 3.21). But once probabilities take irrational values, we have construct our models abstractly.

3.2. Independent trials

In this section we introduce the model of a sequence of independent trials. The simplest kind of trial has two outcomes that we may call *success* and *failure*. Success comes with probability p and failure with probability $1-p$ where $p \in [0,1]$ is a fixed parameter of the model. In this context we introduce six basic probability distributions: Bernoulli, binomial, geometric, negative binomial, Poisson, and exponential. At the end of the section we generalize to trials with more than two outcomes and introduce the multinomial distribution. These distributions will appear in examples and exercises throughout the remainder of the book.

In Example 1.13 we introduced the probability space for a sequence of independent fair coin flips with state space

$$\Omega = \{\omega = (s_i)_{i\in\mathbb{Z}_{>0}} : \text{each } s_i \text{ equals } 0 \text{ or } 1\}. \tag{3.12}$$

In the model 1 represented tails and 0 was heads. We can also use the same model for a sequence of repeated independent experiments with two outcomes: 1 for success and 0 for failure, where the probability of success is $\frac{1}{2}$.

We now generalize this model to allow for a success probability $p \in [0,1]$ for the repeated trial. We use the same sample space Ω given in (1.10), with s_i representing the outcome of the ith trial. To identify the probability measure P we use the same method as before: for a given n-tuple $\mathbf{t} = (t_1, \dots, t_n) \in \{0,1\}^n$ we set the probability of the event

$$A_{n,\mathbf{t}} = \{\omega = (s_k)_{1\leq k<\infty} \in \Omega : (s_1, \dots, s_n) = (t_1, \dots, t_n)\}$$

to be the same as the probability of seeing the outcomes $(t_1, t_2, \dots, t_n)$ as the outcomes of n independent trials with common success probability p. This probability can be computed using the independence property. For a given $1 \le i \le n$ the probability of $\{t_i = 1\}$ is p, and the probability of $\{t_i = 0\}$ is $1-p$, so by independence we should have

$$P(A_{n,\mathbf{t}}) = p^{\sum_{i=1}^n t_i}(1-p)^{n-\sum_{i=1}^n t_i}. \tag{3.13}$$

Note that the sums $\sum_{i=1}^n t_i$ and $n - \sum_{i=1}^n t_i = \sum_{i=1}^n (1-t_i)$ count the number of successes and number of failures in the sequence $(t_1, \dots, t_n)$.

As already explained earlier, without measure theory we cannot prove that this formula determines a unique probability measure P on a σ-algebra that contains all the events of the type that appear in (3.13). However, this shortcoming will not prevent us from doing any of the calculations we wish to tackle.

On the sequence space Ω we introduce random variables $X_1, X_2, X_3, \dots$ that record the outcomes of the successive trials: for a sample point $\omega = (s_i)_{i=1}^\infty$ and a positive integer k, $X_k(\omega) = s_k$. In other words,

$$X_k = \begin{cases} 1, & \text{if trial } k \text{ is a success} \\ 0, & \text{if trial } k \text{ is a failure.} \end{cases} \tag{3.14}$$

Mathematically, the function X_k maps the sequence ω into its kth coordinate s_k. Random variables of this type can be defined on any sequence space. They are called *projection mappings* and *coordinate random variables*.

All events concerning the trials can now be expressed in terms of the random variables $\{X_k\}$. In particular, identity (3.13) takes the following form, for any n-tuple $(t_1, \dots, t_n) \in \{0,1\}^n$:

$$\begin{aligned} P\{X_1 = t_1,\, X_2 = t_2, \dots, X_n = t_n\} &= p^{\sum_{i=1}^n t_i}(1-p)^{\sum_{i=1}^n (1-t_i)} \\ &= \prod_{i=1}^n \bigl(p^{t_i} \cdot (1-p)^{1-t_i}\bigr). \end{aligned} \tag{3.15}$$

Theorem 3.18. *The random variables $X_k, k \ge 1$ are independent, and they all have Bernoulli distributions with parameter p.*

Proof. This is is immediate from Theorem 3.6. The function

$$g(t) = p^t(1-p)^{1-t}, \qquad t \in \{0,1\}$$

is the probability mass function of the Bernoulli distribution with parameter p.

Fix $n \ge 1$. According to (3.15) the joint probability mass function of $X_1, \dots, X_n$ evaluated at $t_1, t_2, \dots, t_n$ with $t_i \in \{0,1\}$ is equal to $\prod_{i=1}^n g(t_i)$. By Theorem 3.6 this means that $X_1, \dots, X_n$ are independent and their marginal probablity mass functions are all equal to g. □

A fundamental fact about independent trials is that, except in the trivial cases $p = 0$ and $p = 1$, *any particular pattern of zeros and ones appears eventually.* We have seen a version of this statement for repeated die rolls in Example 1.22. The proof for repeated trials is essentially the same. (See Exercise 3.3.)

Theorem 3.19. *Assume $0 < p < 1$. Fix a positive integer k and a k-vector $\mathbf{t} = (t_1, \dots, t_k) \in \{0,1\}^k$ of zeros and ones. Then*

$$P\biggl(\bigcup_{n=1}^{\infty}\{(X_n, X_{n+1}, \dots, X_{n+k-1}) = (t_1, \dots, t_k)\}\biggr) = 1. \tag{3.16}$$

The stochastic model of independent trials with two outcomes is now in place. We can turn to defining important probability distributions. For each distribution we explain how it arises from independent trials, give a formal definition, and illustrate it with a numerical example.

Bernoulli distribution. We have introduced the Bernoulli distribution in Definition 2.15. The coordinate random variables $\{X_k\}$ introduced above are Bernoulli random variables with parameter p. Here is a different example.

Example 3.20. Turn over the top card of a well-shuffled standard deck of 52 cards. Let $X = 1$ if the card is an ace, and otherwise $X = 0$. There are four aces in a deck. If we assume that each card is equally likely to be on top, then $P(X = 1) = \frac{4}{52} = \frac{1}{13}$. Thus $X \sim \text{Ber}(\frac{1}{13})$. △

Binomial distribution. The binomial distribution arises from counting successes. Let $S_n = X_1 + X_2 + \cdots + X_n$ be the number of successes in the first n trials. There are $\binom{n}{k}$ n-tuples with exactly k ones and $n-k$ zeros. From (3.13), each particular outcome of n trials with exactly k successes has probability $p^k(1-p)^{n-k}$. Thus by adding up the probabilities of all the different possibilities,

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

This is an example of a binomial random variable with parameters n and p.

Definition 3.21. Let n be a positive integer and $0 \le p \le 1$. A random variable X has the **binomial distribution** with parameters n and p if the possible values of X are $\{0, 1, \dots, n\}$ and the probabilities are

$$P(X = k) = \binom{n}{k}p^k(1-p)^{n-k} \quad \text{for } k = 0, 1, \dots, n.$$

Abbreviate this by $X \sim \text{Bin}(n, p)$. △

The fact that binomial probabilities add to 1 is a particular case of the binomial theorem:

$$\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p + 1 - p)^n = 1. \tag{3.17}$$

Example 3.22. What is the probability that five rolls of a fair die yield two or three sixes?

The repeated trial is a roll of the die, and success means rolling a six. Let Y be the number of sixes that appear in five rolls. Then $Y \sim \text{Bin}(5, \frac{1}{6})$.

$$\begin{aligned} P(Y \in \{2,3\}) &= P(Y = 2) + P(Y = 3) \\ &= \binom{5}{2}(\tfrac{1}{6})^2(\tfrac{5}{6})^3 + \binom{5}{3}(\tfrac{1}{6})^3(\tfrac{5}{6})^2 = \frac{1500}{7776} \approx 0.193. \end{aligned}$$

△

Geometric distribution. Let N be the number of trials needed to see the first success in a sequence of independent trials with success probability p. The mathematical definition of N as a function on Ω is

$$N(\omega) = \inf\{n \in \mathbb{Z}_{>0} : X_n(\omega) = 1\}. \tag{3.18}$$

To find the probability mass function of N, we express the event $\{N = k\}$ in terms of the trial outcome variables and then use independence. For any positive integer k

$$P(N = k) = P(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1-p)^{k-1}p. \tag{3.19}$$

This random variable has the geometric distribution with parameter p.

Definition 3.23. Let $0 < p \le 1$. A random variable X has the **geometric distribution** with success parameter p if the set of possible values of X is $\mathbb{Z}_{>0}$ and X satisfies $P(X = k) = (1-p)^{k-1}p$ for positive integers k. Abbreviate this by $X \sim \mathrm{Geom}(p)$. △

If $p = 0$ there is never a first success and $P(X = k) = 0$ for all positive integers k. So this case does not produce an integer-valued random variable.

Example 3.24. What is the probability that it takes more than seven rolls of a fair die to roll a six?

Let M be the number of rolls of a fair die until the first six. Then $M \sim \mathrm{Geom}(1/6)$.

$$P(M > 7) = \sum_{k=8}^{\infty} P(M = k) = \sum_{k=8}^{\infty} \left(\tfrac{5}{6}\right)^{k-1}\tfrac{1}{6} = \frac{\frac{1}{6}\left(\frac{5}{6}\right)^7}{1 - \frac{5}{6}} = \left(\tfrac{5}{6}\right)^7.$$

△

Example 3.25. The random variables $\{X_k\}$ that give the outcomes of the trials each have probability distribution $P(X_k = 1) = p$ and $P(X_k = 0) = 1 - p$. But what if we choose a trial with a random index N and ask for the distribution of X_N? The distribution can change. For example, if N is the index of the first success as in (3.18) above, then $X_N(\omega) = 1$ for all ω except for the one exceptional ω that has no successes. Hence $P(X_N = 1) = 1$. △

Sometimes the geometric random variable is defined as the number of failures before the first success. This is a shifted version of the definition above. That is, if $X \sim \mathrm{Geom}(p)$ is the number of trials needed for the first success then the number of failures preceding the first success is $Y = X - 1$. The set of possible values of Y is $\mathbb{Z}_{\ge 0}$ and $P(Y = k) = (1-p)^k p$.

Negative binomial distribution. We generalize the geometric distribution by waiting until we see more than one success. Fix a positive integer k. Let L denote the number of trials needed for k successes. The event $L = n$ happens if and only if the first $n-1$ trials have $k-1$ successes and then trial n is a success. The event concerning the first $n-1$ trials is independent of the nth trial. Thus we get

$$\begin{aligned} P(L=n) &= P(S_{n-1}=k-1,\, X_n=1) = P(S_{n-1}=k-1)\, P(X_n=1) \\ &= \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}\cdot p = \binom{n-1}{k-1} p^{k}(1-p)^{n-k}. \end{aligned} \tag{3.20}$$

Definition 3.26. Let k be a positive integer and $0 < p \le 1$. A random variable X has the **negative binomial distribution** with parameters (k,p) if the set of possible values of X is the set of integers $\mathbb{Z}_{\ge k} = \{k, k+1, k+2, \dots\}$ and the probability mass function is

$$P(X=n) = \binom{n-1}{k-1} p^{k}(1-p)^{n-k} \qquad \text{for } n \in \mathbb{Z}_{\ge k}. \tag{3.21}$$

Abbreviate this by $X \sim \text{Negbin}(k,p)$. △

The $\text{Negbin}(1,p)$ distribution is the same as the $\text{Geom}(p)$ distribution.

Remark 3.27. Do the probabilities in (3.21) sum to one? Since these probabilities come from the calculation in (3.20), the only way they can fail to sum to one is that there is positive probability of never seeing the kth success. However, according to Theorem 3.19, with probability one a sequence of k consecutive successes appears eventually and hence in particular, a kth success.

Exercise 3.17 points the way to verifying analytically that

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

△

Example 3.28. What is the probability that it takes exactly seven rolls of a fair die to see three sixes?

Let K be the number of rolls of a fair die until the third six. Then $K \sim \text{Negbin}(3, \frac{1}{6})$.

$$P(K=7) = \binom{6}{2} \left(\tfrac{1}{6}\right)^3 \left(\tfrac{5}{6}\right)^4 = \frac{5^5}{2\cdot 6^6} \approx 0.0335.$$

△

Fact 3.29. Consider a sequence of independent trials with success probability $0 < p \le 1$. Let N_k be the number of trials needed for the kth success. Set $Y_1 = N_1$ and $Y_k = N_k - N_{k-1}$ for $k \ge 2$. Then the random variables $Y_1, Y_2, Y_3, \dots$ are i.i.d. In particular, $N_k - N_{k-1} \sim \text{Geom}(p)$ for each $k \ge 2$.

Proof. We know that $Y_1 = N_1 \sim \text{Geom}(p)$. We need to show that for any fixed $n \ge 2$ the random variables $Y_1, Y_2, \dots, Y_n$ are independent and each has distribution $\text{Geom}(p)$. For this we need to check that for any sequence of positive integers $a_1, a_2, \dots, a_n$ we have

$$P(Y_1 = a_1, Y_2 = a_2, \dots, Y_n = a_n) = \prod_{i=1}^{n} p(1-p)^{a_i - 1}. \tag{3.22}$$

This identity implies the claims. By summing over $a_1, \dots, a_{n-1} \in \mathbb{Z}_{>0}$ we find that $P(Y_n = a_n) = p(1-p)^{a_n - 1}$. This tells that $Y_n \sim \operatorname{Geom}(p)$ for each n. After this, the identity can be re-expressed as

$$P(Y_1 = a_1, Y_2 = a_2, \dots, Y_n = a_n) = \prod_{i=1}^{n} P(Y_i = a_i)$$

which implies the independence. (See Exercise 3.5 for a general version of this reasoning.)

Identity (3.22) follows by noting that the event $\{Y_1 = a_1, Y_2 = a_2, \dots, Y_n = a_n\}$ can be described by a specific outcome of the first $a_1 + \dots + a_n$ trials. Indeed, this is the event that among the first $a_1 + \dots + a_n$ trials there are exactly n successes, and these successes are at positions a_1, $a_1 + a_2, \dots, a_1 + a_2 + \dots + a_n$. The probability of this event is $p^n (1-p)^{\sum_{i=1}^n a_i - n}$, which equals the right-hand side of (3.22). □

Poisson distribution. The next two distributions to arise from independent trials, the Poisson and the exponential, require a special kind of limit. Imagine a sequence of experiments indexed by positive integers $n = 1, 2, 3, \dots$. In the nth round a trial is repeated n times and the number S_n of successes is recorded. If the success probability p is kept fixed, then it can be shown that

$$\lim_{n\to\infty} P(S_n \ge k) = 1 \quad \text{for each positive integer } k.$$

(See Exercise 3.13.) In other words, S_n simply blows up. To get a nontrivial limit, we shrink the success probability p as n grows.

We begin below with the definition of the Poisson distribution. Then we state and prove the Poisson limit of the binomial, and describe the significance of this limit for applications. We close with examples.

Definition 3.30. Let $\lambda > 0$. A random variable X has the **Poisson distribution** with parameter λ if the possible values of X are the nonnegative integers and the probability mass function is

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad \text{for } k \in \{0, 1, 2, \dots\}. \tag{3.23}$$

Abbreviate this by $X \sim \operatorname{Poisson}(\lambda)$. △

Theorem 3.31. *Fix $\lambda > 0$. For positive integers n for which $\lambda/n < 1$, let $S_n \sim \operatorname{Bin}(n, \lambda/n)$. Then*

$$\lim_{n\to\infty} P(S_n = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad \textit{for all} \quad k \in \mathbb{Z}_{\ge 0}. \tag{3.24}$$

Proof. The proof to follow uses the limit $(1 + x/n)^n \to e^x$ as $n \to \infty$, stated as (C.13) in Appendix C.

We rearrange the binomial probability and then observe the limits of the different parts of the expression. Note that k is kept fixed as n is taken to infinity.

$$\begin{aligned}
P(S_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{k!}\,\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \\
&= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right]\frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \\
&\underset{n\to\infty}{\longrightarrow} \frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1\cdot 1 \;=\; e^{-\lambda}\frac{\lambda^k}{k!}. \qquad \square
\end{aligned}$$

Limit (3.24) is called the *law of rare events* because as n grows, the shrinking success probability λ/n means that successes become very rare. This limit suggests that whenever successes are rare in a sequence of independent trials, then the number of successes is well approximated by a Poisson random variable.

The shortcoming of Theorem 3.31 is that it does not give a quantitative error bound for the approximation of a binomial probability with a Poisson probability. Later in Chapter 9 we reprove this limit with a rigorous error bound. For now, we illustrate this approximation with examples.

Example 3.32. Compare $S \sim \text{Bin}(10, \frac{1}{10})$ and $X \sim \text{Poisson}(1)$. We have fairly close agreement, even though $n = 10$ is not very large and the success probability $p = \frac{1}{10}$ is not terribly small. For example,

$$P(S=0) = (\tfrac{9}{10})^{10} \approx 0.3487 \quad \text{while} \quad P(X=0) = e^{-1} \approx 0.3679$$

and

$$P(S=2) = \tfrac{1}{2}\cdot(\tfrac{9}{10})^{9} \approx 0.1937 \quad \text{while} \quad P(X=2) = \tfrac{1}{2}e^{-1} \approx 0.1839.$$

The relative errors are about 5%.

Once we take a value beyond n, the binomial probability has vanished but the Poisson distribution returns a tiny positive probability: for example,

$$P(S=11) = 0 \quad \text{while} \quad P(X=11) = \frac{e^{-1}}{11!} \approx 9\cdot 10^{-9}.$$

$\triangle$

Example 3.33. A delivery company keeps records of accidents on a weekly basis. According to the data, about one out of three weeks is accident-free. Estimate the probability that at least two accidents happen next week.

Let X denote the number of accidents next week. Suppose there is good reason to believe that accidents are rare and happen more or less independently of each other. For example, we can imagine that every time a delivery is made, there is a small chance of an accident. These are the trials with rare successes.

Let us then assume that X is Poisson distributed. We need to estimate its parameter λ. The data suggests the value $P(X=0) = 1/3$, while the Poisson formula is $P(X=0) = e^{-\lambda}$. To match these take $\lambda = \log 3$. (Note: log is the natural logarithm with base e, often also denoted by ln.) The probability estimate is

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) = 1 - e^{-\lambda} - e^{-\lambda}\lambda \\ &= 1 - \tfrac{1}{3} - \tfrac{1}{3}\log 3 \approx 0.3. \end{aligned}$$

△

Exponential distribution.

Definition 3.34. Let $0 < \lambda < \infty$. A random variable X has the **exponential distribution** with parameter λ if X has density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{3.25}$$

on the real line. Abbreviate this by $X \sim \text{Exp}(\lambda)$.

The $\text{Exp}(\lambda)$ distribution is also called the *exponential distribution with rate* λ. The cumulative distribution function of the $\text{Exp}(\lambda)$ distribution is given by

$$F(t) = \int_{-\infty}^{t} f(x)dx = \int_{0}^{t} \lambda e^{-\lambda x}\, dx = 1 - e^{-\lambda t} \qquad \text{for } t \geq 0 \tag{3.26}$$

and $F(t) = 0$ for $t < 0$. By letting $t \to \infty$ in (3.26) we see that $\int_{-\infty}^{\infty} f = 1$ so f is indeed a probability density function. It is often useful that an exponential random variable can be characterized by its tail probability: $X \sim \text{Exp}(\lambda)$ if and only if $P(X > t) = e^{-\lambda t}$ for all $t > 0$.

The exponential distribution often models a waiting time. Since in the probability $e^{-\lambda t}$ the quantity λt should be dimensionless, the natural unit for the rate λ is 1/time.

Uniquely among absolutely continuous distributions, the exponential distribution possesses the so-called *memoryless property* stated in the next theorem.

Theorem 3.35. *Suppose that* $X \sim \text{Exp}(\lambda)$*. Then for any* $s, t > 0$*,*

$$P(X > t + s \,|\, X > t) = P(X > s). \tag{3.27}$$

Proof. From the definition of conditional probability,

$$\begin{aligned} P(X > t + s \,|\, X > t) &= \frac{P(X > t + s,\, X > t)}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s). \end{aligned}$$

□

The memoryless property means that if $X \sim \text{Exp}(\lambda)$ then the random variable $X - t$ has the same distribution as X **given that** $X > t$. Thus if X models the lifetime of an appliance, then if the appliance is in operation at time t then the rest of its lifetime will have the same distribution as a brand new appliance.

Example 3.36. Assume that the first phone call of the day to a help center arrives after an exponentially distributed random time T with rate 1/10 when time is measured in minutes. Suppose 15 minutes have gone by without the first call. What is the probability that the call arrives in the next five minutes?

By assumption $T \sim \mathrm{Exp}(\frac{1}{10})$. By the memoryless property,

$$P(T > 20 \,|\, T > 15) = P(T > 5) = e^{-5/10} = e^{-1/2}.$$

Thus the first call arrives in the next 5 minutes with probability $1 - e^{-1/2} \approx 0.39$. △

As the last item we show that the exponential distribution arises from independent trials. We use the same scaled success probability that produced the Poisson distribution, and this time look at the time until the first success in an altered time scale. In the nth round consider an infinite sequence of independent trials with success probability λ/n. $\lambda > 0$ is fixed and n is taken large enough so that $\lambda/n < 1$. The new feature is that the kth trial takes place at time k/n. Let T_n denote the time of the first success in this speeded up trials process. Then

$$P\big(T_n = \tfrac{k}{n}\big) = \big(1 - \tfrac{\lambda}{n}\big)^{k-1} \tfrac{\lambda}{n} \qquad \text{for } k \geq 1. \tag{3.28}$$

Equivalently, $nT_n \sim \mathrm{Geom}(\lambda/n)$. The theorem is that the probability distribution of T_n converges to the exponential distribution. For computational convenience we take the limit below in terms of the tail probabilities.

Theorem 3.37. *Fix $\lambda > 0$. Consider n large enough so that $\lambda/n < 1$. Suppose that for each n, the random variable T_n satisfies $nT_n \sim \mathrm{Geom}(\lambda/n)$. Then*

$$\lim_{n\to\infty} P(T_n > t) = e^{-\lambda t} \qquad \text{for all real } t \geq 0. \tag{3.29}$$

Proof. Observe first that since $P(T_n > 0) = 1$, limit (3.29) holds for $t = 0$. We can assume $t > 0$ for the rest of the proof.

For any real x, the floor function

$$\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \leq x\} \tag{3.30}$$

gives the largest integer less than or equal to x. An integer k satisfies $k > x$ if and only if $k \geq \lfloor x \rfloor + 1$. (For example, an integer k satisfies $k > 3.7$ if and only if $k \geq 3 + 1$.) We compute the tail probability of T_n for $t > 0$:

$$\begin{aligned} P\big(T_n > t\big) = P\big(nT_n > nt\big) &= \sum_{k:k>nt} \big(1 - \tfrac{\lambda}{n}\big)^{k-1} \tfrac{\lambda}{n} = \sum_{k=\lfloor nt \rfloor+1}^{\infty} \big(1 - \tfrac{\lambda}{n}\big)^{k-1} \tfrac{\lambda}{n} \\ &= \big(1 - \tfrac{\lambda}{n}\big)^{\lfloor nt \rfloor} = \big(1 - \tfrac{\lambda}{n}\big)^{nt} \big(1 - \tfrac{\lambda}{n}\big)^{\lfloor nt \rfloor - nt} \\ &= \big(1 - \tfrac{\lambda t}{nt}\big)^{nt} \frac{1}{\big(1 - \frac{\lambda}{n}\big)^{nt - \lfloor nt \rfloor}}. \end{aligned} \tag{3.31}$$

The assumption $t > 0$ was used at the end to put t in the denominator.

Since $0 \leq nt - \lfloor nt \rfloor \leq 1$ for all n and t, we see that $\big(1 - \frac{\lambda}{n}\big)^{nt - \lfloor nt \rfloor} \to 1$ as $n \to \infty$. The expression $\big(1 - \frac{\lambda t}{nt}\big)^{nt}$, and hence the desired probability $P(T_n > t)$, converges to $e^{-\lambda t}$ by the limit (C.13). □

Remark 3.38. Theorems 3.31 and 3.37 are examples of *limits in distribution*. The key feature of these limits is that it is the probability distributions that converge, and not the values of the random variables themselves. Such limits will be defined precisely in Chapter 6. △

Multinomial distribution. Consider the problem of counting the occurrences of different outcomes in repeated trials with more than two outcomes. The setting is the following. A trial has r possible outcomes labeled $1, \dots, r$. In each trial outcome j appears with probability p_j, and $p_1 + p_2 + \cdots + p_r = 1$. Perform n independent repetitions of this trial. For $j = 1, \dots, r$ let X_j denote the number of times outcome j appears among the n trials. Then the joint distribution of $(X_1, \dots, X_r)$ is the *multinomial distribution* defined below.

Definition 3.39. Let n and r be positive integers and let $p_1, p_2, \dots, p_r$ be positive reals such that $p_1 + p_2 + \cdots + p_r = 1$. The random vector $\mathbf{X} = (X_1, \dots, X_r)$ has the **multinomial distribution** with parameters n, r and $p_1, \dots, p_r$ if the possible values of $\mathbf{X}$ are integer vectors $(k_1, \dots, k_r)$ such that $k_j \geq 0$ and $k_1 + \cdots + k_r = n$, and the joint probability mass function is given by

$$P\big(X_1 = k_1, X_2 = k_2, \dots, X_r = k_r\big) = \binom{n}{k_1, k_2, \dots, k_r} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r} \tag{3.32}$$

where the **multinomial coefficient** is defined by

$$\binom{n}{k_1, k_2, \dots, k_r} = \frac{n!}{k_1!\, k_2! \, \cdots \, k_r!}.$$

Abbreviate this by $(X_1, \dots, X_r) \sim \text{Mult}(n, r, p_1, \dots, p_r)$. △

The justification for the claim that this is the correct joint distribution for the variables $(X_1, \dots, X_r)$ that count the occurrences of outcomes $1, \dots, r$ in n repeated trials lies on two observations. Consider nonnegative integers $k_1, \dots, k_r$ such that $k_1 + \cdots + k_r = n$.

(i) By the independence of the trials, any particular sequence of outcomes from the n trials that yields outcome j exactly k_j times has probability $p_1^{k_1} \cdots p_r^{k_r}$.

(ii) The multinomial coefficient $\binom{n}{k_1, k_2, \dots, k_r}$ counts the number of ways n items can be labeled with integers $1, \dots, r$ so that label j appears exactly k_j times. This is exactly the number of outcomes from the n trials contained in the event $\{X_1 = k_1, X_2 = k_2, \dots, X_r = k_r\}$.

Summing over the equally likely arrangements gives the formula

$$P\big(X_1 = k_1, X_2 = k_2, \dots, X_r = k_r\big) = \binom{n}{k_1, k_2, \dots, k_r} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}.$$

The special case with $r = 2$ is the binomial distribution. In that case $X_2 = n - X_1$ so only one of the variables is needed.

Example 3.40. Suppose an urn contains 1 green, 2 red and 3 yellow balls. We sample a ball with replacement 10 times. Find the probability that green appeared 3 times, red twice, and yellow 5 times.

Let X_G, X_R and X_Y denote the number of green, red and yellow balls in the sample. Then $(X_G, X_R, X_Y) \sim \text{Mult}(10, 3, \frac{1}{6}, \frac{2}{6}, \frac{3}{6})$ and

$$\begin{aligned} &P(\text{green appeared 3 times, red twice, and yellow 5 times}) \\ &\qquad = P(X_G = 3, X_R = 2, X_Y = 5) = \frac{10!}{3!\,2!\,5!}\left(\frac{1}{6}\right)^3 \left(\frac{2}{6}\right)^2 \left(\frac{3}{6}\right)^5 \approx 0.0405. \end{aligned}$$

△

The next two examples look at marginal distributions of a multinomial.

Example 3.41. Let $(X_1, \dots, X_r) \sim \text{Mult}(n, r, p_1, \dots, p_r)$. Find the distribution of X_1.

The random variable X_1 counts the number of times that outcome 1 appears among the n trials. Considering outcome 1 as a success and any other outcome as a failure, we see that X_1 is a binomial random variable with parameters n and p_1. △

Example 3.42. We roll a die 100 times. Find the probability that among the 100 rolls we observe exactly 22 ones and 17 fives.

Denote by X_i the number of times i appears among the 100 rolls. Then $(X_1, \dots, X_6)$ has multinomial distribution with $n = 100$, $r = 6$ and $p_1 = \cdots = p_6 = \frac{1}{6}$. We need to compute $P(X_1 = 22, X_5 = 17)$. This can be computed from the joint probability mass function of $(X_1, \dots, X_6)$ by summing over the possible values of X_2, X_3, X_4 and X_6. But there is a simpler way.

Since we are only interested in the rolls that are 1 or 5, we can combine the other outcomes into a new outcome: let $Y = X_2 + X_3 + X_4 + X_6$ denote the number of times we rolled something other than 1 or 5. The probability that a particular roll is 2, 3, 4, or 6 is $\frac{2}{3}$, and thus

$$(X_1, X_5, Y) \sim \text{ Mult}(100, 3, \tfrac{1}{6}, \tfrac{1}{6}, \tfrac{2}{3}).$$

Now we can express the required probability using the multinomial joint probability mass function:

$$\begin{aligned} P(X_1 = 22, X_5 = 17) &= P(X_1 = 22, X_5 = 17, Y = 61) \\ &= \frac{100!}{22!\,17!\,61!} \left(\tfrac{1}{6}\right)^{22} \left(\tfrac{1}{6}\right)^{17} \left(\tfrac{2}{3}\right)^{61} \approx 0.0037. \end{aligned}$$

△

3.3. Convolution

Of the various operations performed on independent random variables, addition is especially frequent. Its analysis introduces us to the *convolution*, a notion that appears in various parts of mathematics.

As a mathematical concept, convolution is a way of multiplying functions to produce new functions. The operation is denoted by $*$. The convolution of two functions on the real line is the function $f * g$ whose value at x is defined by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(y) g(x - y)\, dy, \tag{3.33}$$

provided of course that these integrals are well-defined for all $x \in \mathbb{R}$. Change of variable from y to $x - y$ shows that convolution is symmetric: $f * g = g * f$.

The discrete version of convolution operates on sequences indexed by $\mathbb{Z}$. The convolution product $\mathbf{a} * \mathbf{b}$ of two sequences $\mathbf{a} = \{a_k\}$ and $\mathbf{b} = \{b_k\}$ is the new sequence whose nth entry is defined by

$$(\mathbf{a} * \mathbf{b})_n = \sum_{k\in\mathbb{Z}} a_k b_{n-k}, \quad n \in \mathbb{Z}. \tag{3.34}$$

To find a probabilistic meaning for convolution we look at the distribution of $X+Y$ for two independent random variables X and Y.

Discrete random variables. If X and Y are discrete, then so is $X+Y$. This is clear if both X and Y only takes finitely many values, as in this case $X+Y$ can only take finitely many values as well. In the more general case, when X and/or Y takes countably many values one can use basic set theory to show that $X+Y$ can only take countably many values, too.

Since $X+Y$ is discrete, we can describe its distribution with its probability mass function. For a given a we can compute $P(X+Y=z)$ by decomposing $\{X+Y=z\}$ as the disjoint union of events of the form $\{X=x, Y=z-x\}$ where x is a possible value of X and $z-x$ is a possible value of Y. Then using independence we get

$$\begin{aligned} P(X+Y=z) &= \sum_x P(X=x, X+Y=z) = \sum_x P(X=x, Y=z-x) \\ &= \sum_x P(X=x)P(Y=z-x). \end{aligned}$$

When both X and Y take only integer values then the same is true for $X+Y$, and we can express the previous formula as

$$P(X+Y=n) = \sum_{k=-\infty}^{\infty} P(X=k)P(Y=n-k). \tag{3.35}$$

The formula (3.35) describes how to obtain the probability mass function of $X+Y$ in terms of the probability mass functions of X and Y. The operation on the right is called convolution, and it is denoted by the symbol $*$. Thus we can write (3.35) as $p_{X+Y} = p_X * p_Y$. We state this as the next theorem.

Theorem 3.43. *If X and Y are independent $\mathbb{Z}$-valued random variables with probability mass functions p_X and p_Y, then the probability mass function of $X+Y$ is*

$$p_{X+Y}(n) = p_X * p_Y(n) = \sum_{k\in\mathbb{Z}} p_X(k)\, p_Y(n-k) = \sum_{\ell\in\mathbb{Z}} p_X(n-\ell)\, p_Y(\ell). \tag{3.36}$$

If both X and Y take only nonnegative integer values, then the same is true for $X+Y$. For a given $n \geq 0$ integer k and $n-k$ are both nonnegative if and only if $0 \leq k \leq n$. Hence in this case (3.35) gives

$$P(X+Y=n) = \sum_{k=0}^{n} P(X=k)P(Y=n-k). \tag{3.37}$$

Example 3.44. Suppose that X and Y are independent with $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. Find the distribution of $X+Y$.

Since X and Y take only nonnegative integer values, the same will hold for $X+Y$ as well and we may use the convolution formula (3.37). For $n \ge 0$ we get

$$\begin{aligned} P(X+Y=n) &= \sum_{k=0}^{n} P(X=k)P(Y=n-k) \\ &= \sum_{k=0}^{n} \frac{\lambda^k}{k!}e^{-\lambda}\frac{\mu^{n-k}}{(n-k)!}e^{-\mu} = e^{-(\lambda+\mu)}\sum_{k=0}^{n}\frac{\lambda^k}{k!}\frac{\mu^{n-k}}{(n-k)!}. \end{aligned}$$

We can evaluate the sum by noting that $\frac{1}{k!(n-k)!} = \frac{1}{n!}\binom{n}{k}$ and hence

$$\sum_{k=0}^{n}\frac{\lambda^k}{k!}\frac{\mu^{n-k}}{(n-k)!} = \frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda^k\mu^{n-k} = \frac{(\lambda+\mu)^n}{n!}$$

by the binomial theorem. This leads to

$$P(X+Y=n) = \frac{(\lambda+\mu)^n}{n!}e^{-(\lambda+\mu)},$$

which means that $X+Y \sim \text{Poisson}(\lambda+\mu)$.

△

Example 3.45. Suppose that X, Y are independent Geom(p) distributed random variables. Find the distribution of $X+Y$.

Since X and Y take only positive integer values, $X+Y$ will be a positive integer that is at least 2. For $n \ge 2$ we get

$$\begin{aligned} P(X+Y=n) &= \sum_{k=-\infty}^{\infty} P(X=k)P(Y=n-k) = \sum_{k=1}^{n-1} P(X=k)P(Y=n-k) \\ &= \sum_{k=1}^{n-1} p(1-p)^{k-1}p(1-p)^{n-k-1} = \sum_{k=1}^{n-1} p^2(1-p)^{n-2} \\ &= (n-1)p^2(1-p)^{n-2}. \end{aligned}$$

We restricted the summation in the convolution formula to $2 \le k \le n-1$ because these are the indices where both k and $n-k$ are positive integers. Note that the resulting probability mass function is the same as (3.21) with $k=2$, hence $X+Y \sim$ Negbin$(2,p)$.

We can give a simple explanation for our result using Example 3.29. Consider a sequence of independent trials with success probability $0 < p \le 1$. Let X denote the number of trials needed for the first success and let N denote the number of trials needed for the second success. We have seen in Section 3.2 that $X \sim$ Geom(p) and $N \sim$ Negbin$(2,p)$. On the other hand, by Example 3.29 the random variables X and $N-X$ are independent and identically distributed. Thus the sum of two independent Geom(p) distributed random variables ($N = X + (N-X)$) has negative binomial distribution with parameters $(2,p)$. △

Continuous random variables. Suppose now that X and Y are absolutely continuous and independent. If the probability density functions are f_X and f_Y then the joint probability density function of X, Y is $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. For a given

$z \in \mathbb{R}$ we compute the probability $P(X+Y \le z)$ by integrating the joint density over the region $\{(x,y) : x+y \le z\}$.

$$
\begin{aligned}
P(X+Y \le z) &= \iint_{x+y\le z} f_X(x) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} f_X(x) \Big(\int_{-\infty}^{z-x} f_Y(y)\, dy \Big)\, dx \\
&= \int_{-\infty}^{\infty} f_X(x) \Big(\int_{-\infty}^{z} f_Y(w-x)\, dw \Big)\, dx \\
&= \int_{-\infty}^{z} \Big(\int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx \Big)\, dw
\end{aligned}
$$

In the third line we introduced the new variable $w = x+y$ in the dy integral. Then we switched the order of the two integrals. Our computation gives the cumulative distribution function of $X+Y$. Differentiating this with respect to z identifies the probability density function as

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx. \tag{3.38}$$

The identity (3.38) is the continuous version of the discrete convolution formula (3.35). The right side of the equation is called the convolution of f_X and f_Y, and denoted by $f_X * f_Y$.

Theorem 3.46. *If X and Y are independent absolutely continuous random variables with density functions f_X and f_Y then the density function of $X+Y$ is*

$$f_{X+Y}(z) = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx = \int_{-\infty}^{\infty} f_X(z-x) f_Y(x)\, dx. \tag{3.39}$$

Example 3.47. Suppose that X and Y are independent with distributions $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Exp}(\lambda)$. Find the distribution of $X+Y$.

X and Y have continuous distributions with probability densities

$$f_X(x) = f_Y(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0 \\ 0, & x < 0. \end{cases}$$

We use the convolution formula (3.38) to compute the probability density of $X+Y$. If $z < 0$ then either x or $z-x$ is negative, hence $f_{X+Y}(z) = 0$ in this case. (This also follows from the fact that X and Y are nonnegative, hence $X+Y$ must be nonnegative as well.) For $z \ge 0$ we have

$$
\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx = \int_0^z f_X(x) f_Y(z-x) dx \\
&= \int_0^z \lambda e^{-\lambda x} \lambda e^{-\lambda(z-x)} dx = \lambda^2 \int_0^z e^{-\lambda z} dx = \lambda^2 z e^{-\lambda z}.
\end{aligned}
$$

Note that we restricted the integration to $[0,z]$ in the first step because outside this interval $f_X(x) f_Y(z-x) = 0$.

Thus the probability density function of $X + Y$ is given by

$$f_{X+Y}(z) = \begin{cases} \lambda^2 z e^{-\lambda z}, & z \ge 0 \\ 0, & z < 0. \end{cases}$$

$X + Y$ has the gamma distribution with parameters $(2, \lambda)$, defined below. △

The gamma distribution arose in the example above. We give a formal definition. First, the *gamma function* is defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx \qquad \text{for } r > 0. \tag{3.40}$$

The gamma function generalizes the factorial to real values: if n is a positive integer then $\Gamma(n) = (n-1)!$.

Definition 3.48. Let $r, \lambda > 0$. A random variable X has the **gamma distribution** with parameters (r, λ) if X is nonnegative and has probability density function

$$f_X(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} \qquad \text{for } x > 0, \tag{3.41}$$

and $f_X(x) = 0$ for $x \le 0$. We abbreviate this $X \sim \text{Gamma}(r, \lambda)$. △

3.4. Exchangeable random variables

Suppose we deal the entire deck of cards in order. What is the probability that the 17th and 27th cards are both aces? The arrangements could be counted, but shouldn't there be a quicker way? After all, why should the 17th and 27th cards be any more or less likely to be aces than the first two cards, giving the answer $\frac{4 \cdot 3}{52 \cdot 51}$? This is indeed a correct way of reasoning. There is a symmetry among the draws, called exchangeability.

To define exchangeability precisely, we use permutations. A *permutation* of the numbers $\{1, 2, \dots, n\}$ is any bijection $\sigma : \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$. There are several standard ways of writing a permutation.

Example 3.49. Take $n = 5$ and let

$$\sigma(1) = 3,\ \sigma(2) = 5,\ \sigma(3) = 1,\ \sigma(4) = 2,\ \sigma(5) = 4.$$

This same permutation can be expressed also as the list $\sigma = (3, 5, 1, 2, 4)$. The *inverse permutation* σ^{-1} of σ is

$$\sigma^{-1}(1) = 3,\ \sigma^{-1}(2) = 4,\ \sigma^{-1}(3) = 1,\ \sigma^{-1}(4) = 5,\ \sigma^{-1}(5) = 2,$$

or as a list, $\sigma^{-1} = (3, 4, 1, 5, 2)$. As for any bijection and its inverse, $\sigma(i) = j$ if and only if $\sigma^{-1}(j) = i$. △

Definition 3.50. Random variables $X_1, \dots, X_n$ are **exchangeable** if for any permutation σ on $\{1, 2, \dots, n\}$, the joint distribution of $(X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)})$ is the same as the joint distribution of $(X_1, X_2, \dots, X_n)$. In other words, permuting the random variables does not change the joint distribution.

Verifying exchangeability boils down to checking that either the joint cumulative distribution function , the joint probability mass function, or the joint density function is a symmetric function. (A function is symmetric if its value is not altered by permuting its arguments.) These cases are collected in the next theorem.

Theorem 3.51.

(i) *The random variables* $X_1, \dots, X_n$ *are exchangeable if for any permutation* σ *on* $\{1, \dots, n\}$ *and for any choice of real numbers* $x_1, x_2, \dots, x_n$ *we have*

$$P(X_1 \le x_1, \dots, X_n \le x_n) = P(X_1 \le x_{\sigma(1)}, \dots, X_n \le x_{\sigma(n)}).$$

(ii) *Suppose* $X_1, \dots, X_n$ *are discrete. Then exchangeability is equivalent to*

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_{\sigma(1)}, \dots, X_n = x_{\sigma(n)}).$$

for all permutations σ *on* $\{1, \dots, n\}$ *and for all choices of real numbers* $x_1, x_2, \dots, x_n$.

(iii) *Suppose* $X_1, \dots, X_n$ *are jointly absolutely continuous with joint density function* $f_{X_1,\dots,X_n}$. *Then exchangeability is equivalent to having the identity*

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = f_{X_1,\dots,X_n}(x_{\sigma(1)}, \dots, x_{\sigma(n)})$$

for all permutations σ *on* $\{1, \dots, n\}$ *and for all choices of real numbers* $x_1, x_2, \dots, x_n$, *except possibly on a set of zero volume.*

Proof. Part (i). $(X_1, \dots, X_n)$ has the same joint distribution as $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ if the joint cumulative distribution functions agree:

$$P(X_1 \le x_1, \dots, X_n \le x_n) = P(X_{\sigma(1)} \le x_1, \dots, X_{\sigma(n)} \le x_n)$$

for all $x_1, \dots, x_n$. Inside the right-hand probability the list of inequalities

$$X_{\sigma(1)} \le x_1, X_{\sigma(2)} \le x_2, \dots, X_{\sigma(n)} \le x_n$$

can be equivalently represented as

$$X_1 \le x_{\sigma^{-1}(1)}, X_2 \le x_{\sigma^{-1}(2)}, \dots, X_n \le x_{\sigma^{-1}(n)}.$$

The inequalities were simply listed in a different order. This is easy to see through an example: with $n = 3$, $\sigma = (3, 1, 2)$ and $\sigma^{-1} = (2, 3, 1)$, the inequalities $X_3 \le x_1, X_1 \le x_2, X_2 \le x_3$ are the same as $X_1 \le x_2, X_2 \le x_3, X_3 \le x_1$.

We now have the statement that $X_1, \dots, X_n$ are exchangeable if and only if

$$P(X_1 \le x_1, \dots, X_n \le x_n) = P(X_1 \le x_{\sigma^{-1}(1)}, \dots, X_n \le x_{\sigma^{-1}(n)})$$

for all real $x_1, \dots, x_n$ and all permutations σ on $\{1, \dots, n\}$. As σ ranges over all permutations, so does σ^{-1}. Hence the statement above is the same as saying that $X_1, \dots, X_n$ are exchangeable if and only if

$$P(X_1 \le x_1, \dots, X_n \le x_n) = P(X_1 \le x_{\sigma(1)}, \dots, X_n \le x_{\sigma(1)})$$

for all permutations σ on $\{1, \dots, n\}$.

The proof of (ii) and (iii) follow from (i). We leave the details as exercises. □

In this course we deal with the following two special cases of exchangeable random variables, namely (i) i.i.d. and (ii) sampling without replacement.

Example 3.52. (Independent, identically distributed random variables.) If $X_1, X_2, \dots, X_n$ are i.i.d. then they are exchangeable.

Denote the c.d.f. of X_i by F. Then the joint c.d.f. of the random variables is $P(X_1 \le x_1, \dots, X_n \le x_n) = \prod_{i=1}^n F(x_i)$. This product remains the same if we replace $x_1, \dots, x_n$ with a permutation of these numbers. Hence, by Theorem 3.51 the random variables are indeed exchangeable. △

Example 3.53. (Sampling without replacement.) Let A be a finite set of size n. Suppose that $X_1, X_2, \dots, X_n$ is a sample of size n from A without replacement. Then the random variables $X_1, \dots, X_n$ are exchangeable.

We may assume that $A = \{1, 2, \dots, n\}$. The joint probability mass function of $X_1, \dots, X_n$ is

$$P(X_1 = x_1, \dots, X_n = x_n) = \frac{1}{n!}$$

for every permutation $(x_1, \dots, x_n)$ of $(1, 2, \dots, n)$, and zero otherwise. The statement now follows from Theorem 3.51. △

We state two corollaries of the characterization of exchangeability. These results tell us how we can apply exchangeability in practical calculations.

Corollary 3.54 (Producing new exchangeable random variables)**.** *Suppose that $X_1, \dots, X_n$ are exchangeable.*

(i) *If $1 \le m \le n$ then $X_1, \dots, X_m$ are also exchangeable.*

(ii) *Let $g : \mathbb{R} \to \mathbb{R}$ be a function and define $Y_k = g(X_k)$. Then the random variables $Y_1, \dots, Y_n$ are also exchangeable.*

Corollary 3.55 (Applications of exchangeability)**.** *If $X_1, \dots X_n$ are exchangeable and $k_1, \dots, k_m$ are distinct numbers from $\{1, 2, \dots, n\}$ then the following statements hold.*

(i) *$X_{k_1}, \dots, X_{k_m}$ have the same joint distribution as $X_1, \dots, X_m$.*

(ii) *For all Borel sets $B \subset \mathbb{R}^m$,*

$$P((X_{k_1}, \dots, X_{k_m}) \in B) = P((X_1, \dots, X_m) \in B).$$

Example 3.56. We shuffle a deck of cards and deal each of the 52 cards one by one. What is the probability that we get an ace for the 10th card and a king for the 37th?

If we could change the numbers 10 and 37 in the problem to 1 and 2 then the problem can easily be solved with counting favorable outcomes:

$$P(\text{1st card is an ace, 2nd card is a king}) = \frac{4 \cdot 4}{52 \cdot 51} = \frac{4}{663}.$$

How can we justify the following equality?

(3.42)
$$P(\text{1st card is ace, 2nd card is king}) = P(\text{10th card is ace, 37th card is king})$$

Imagine that we encode the 52 cards in the deck with the numbers $1, 2, \dots, 52$. Let X_i be the card dealt in the ith step. Then $X_1, \dots, X_{52}$ is a sample of size 52 without replacement from the set $\{1, \dots, 52\}$, and hence these random variables are exchangeable. Let A be the set of aces and K the set of kings in the deck. Then

$$P(\text{10th card is ace, 37th card is king}) = P(X_{10} \in A, X_{37} \in K),$$

and by the exchangeability of $X_1, \dots, X_{52}$ we have

$$P(X_{10} \in A, X_{37} \in K) = P(X_1 \in A, X_2 \in K) = P(\text{1st card is ace, 2nd card is king}).$$

This shows that (3.42) holds.

Another way to see (3.42) is to encode the values of the cards with $1, 2, \dots, 13$ (with 1 corresponding to ace, and 13 corresponding to king), and denoting the function that gives the value of a card by g. Then exchangeability gives

$$\begin{aligned} P(\text{10th card is ace, 37th card is king}) &= P(g(X_{10}) = 1, g(X_{37}) = 13) \\ &= P(g(X_1) = 1, g(X_2) = 13) \\ &= P(\text{1st card is ace, 2nd card is king}). \end{aligned}$$

$\triangle$

Example 3.57. Suppose that $X_1, \dots, X_8$ are i.i.d. random variables that are uniform on $[1, 2]$. What is the probability that X_4 is the largest?

The random variables are absolutely continuous, and because they are independent, they are jointly continuous. (See Theorem 3.11.) Since they have a joint probability density, the probability that two (or more) of these random variables are equal is zero. Hence we can always pick the largest one uniquely, and

$$\sum_{k=1}^{8} P(X_k \text{ is the largest}) = 1.$$

Because the random variables are i.i.d., they are also exchangeable. But that means that $P(X_k \text{ is the largest})$ does not depend on k, and

$$1 = \sum_{k=1}^{8} P(X_k \text{ is the largest}) = 8P(X_1 \text{ is the largest}).$$

That means that $P(X_1 \text{ is the largest}) = \frac{1}{8}$ and $P(X_k \text{ is the largest}) = \frac{1}{8}$ for all $1 \le k \le 8$. $\triangle$

3.5. Simple random walk

We begin with a famous gambling problem as an introduction to simple random walk.

Example 3.58 (Gambler's ruin)**.** You play repeatedly the following gamble. A fair coin is tossed: heads you win a dollar, tails you lose a dollar. You start playing with x dollars in your pocket. You choose a target $M > x$. Then you play until you either reach M dollars or lose all your money. A question of obvious interest: what is the probability that you reach M dollars before going broke?

The first question though is whether the game is sure to end. From our past discussion of independent trials we know that, with probability one, if the coin is flipped long enough there will be a run of M consecutive heads. If the game had not ended by the start of this run, it is certainly over by the time the run is complete, because with any amount of money between 1 and $M-1$ in your pocket, M straight heads would push you up to M or above.

To find the probability of success it turns out useful to consider all the possible initial states x. An equation comes from the following observation. If the first coin flip is heads, we find ourselves at $x+1$. At that point the future of the game is the same as if the game had started at $x+1$. And similarly, if the first flip sends us to $x-1$, we basically restart the game at $x-1$. This reasoning can be put on firmer footing with the law of total probability. Let P_x denote the probability measure of the entire game when the initial position is x. Then, for $0 < x < M$,

$$\begin{aligned} P_x(\text{reach } M \text{ before } 0) &= \tfrac{1}{2}P_x(\text{reach } M \text{ before } 0 \,|\, \text{first flip heads}) \\ &\quad + \tfrac{1}{2}P_x(\text{reach } M \text{ before } 0 \,|\, \text{first flip tails}) \\ &= \tfrac{1}{2}P_{x+1}(\text{reach } M \text{ before } 0) + \tfrac{1}{2}P_{x-1}(\text{reach } M \text{ before } 0). \end{aligned} \tag{3.43}$$

Introduce succinct notation $p_x = P_x(\text{reach } M \text{ before } 0)$ for the probability of reaching M before going broke, when the initial amount of money is x. Natural boundary conditions are $p_0 = 0$ and $p_M = 1$ and (3.43) gives the equation

$$p_x = \tfrac{1}{2}p_{x+1} + \tfrac{1}{2}p_{x-1} \qquad \text{for } 0 < x < M.$$

Rearranging the equation gives

$$p_{x+1} - p_x = p_x - p_{x-1}.$$

This means that $p_0, p_1, \dots, p_M$ is an *arithmetic progression* and we can solve for each p_x. For any $k \in \{1, \dots, M\}$,

$$1 = p_M - p_0 = \sum_{x=1}^{M}(p_x - p_{x-1}) = M(p_k - p_{k-1})$$

from which $p_k - p_{k-1} = 1/M$. Then for any x,

$$p_x = p_x - p_0 = \sum_{k=1}^{x}(p_k - p_{k-1}) = \frac{x}{M}.$$

To summarize, with initial wealth x the probability of leaving the game with M dollars is $\frac{x}{M}$. Exercise 3.42 asks you to find the probability of success when the coin is not fair.

The reasoning in (3.43) from conditioned P_x to $P_{x\pm1}$ is intuitive and not rigorous. It could be made rigorous by considering all the countably many distinct finite sequences of coin flips that constitute the event $\{\text{reach } M \text{ before } 0\}$. However, this is not a worthwhile endeavor except as an exercise in technical power. The whole matter can be handled much more efficiently with a small dose of Markov chain theory covered in any text on stochastic processes. △

The amount of money in your pocket in the previous game changes after each coin flip: it increases by 1 with probability 1/2 and decreases by 1 with probability 1/2, and these successive steps are independent of each other. This random evolution is called a (symmetric) simple random walk. Here is the definition.

Definition 3.59. Fix $p \in (0,1)$. Let $X_1, X_2, X_3, \dots$ be i.i.d. random variables with $P(X_i = 1) = p$ and $P(X_i = -1) = 1 - p$. Let S_0 be an integer. (If S_0 is also random, then S_0 is independent of the random variables $\{X_i\}$.) For $n \ge 1$ define

$$S_n = S_0 + X_1 + X_2 + \cdots + X_n.$$

The random sequence $S_0, S_1, S_2, \dots$ is the **simple random walk** (SRW) with initial position S_0. If an initial position is not specified, then $S_0 = 0$.

If $p = \frac{1}{2}$ then $\{S_n\}$ is called **symmetric simple random walk** (SSRW), while if $p \neq \frac{1}{2}$, then $\{S_n\}$ is **asymmetric simple random walk**. △

The term *stochastic process* means in general an indexed collection of random variables. Definition 3.59 has two stochastic processes: $\{X_i\}_{i \in \mathbb{Z}_{>0}}$ is an example of an i.i.d. process, while the SRW $\{S_n\}_{n \in \mathbb{Z}_{\ge 0}}$ is an example of a *Markov chain*. In both cases the index represents *time*. Figure 1 illustrates a SRW path started at the origin. The integer points (n, S_n) of the path are connected with line segments to produce a continuous, piecewise linear path.

More generally, random walk in $\mathbb{R}^d$ is defined by $S_n = S_0 + \sum_{i=1}^n X_i$ for time indices $n = 0, 1, 2, \dots$ where the $\{X_i\}$ are i.i.d. $\mathbb{R}^d$-valued random vectors and the initial point S_0 is another independent random vector in $\mathbb{R}^d$.

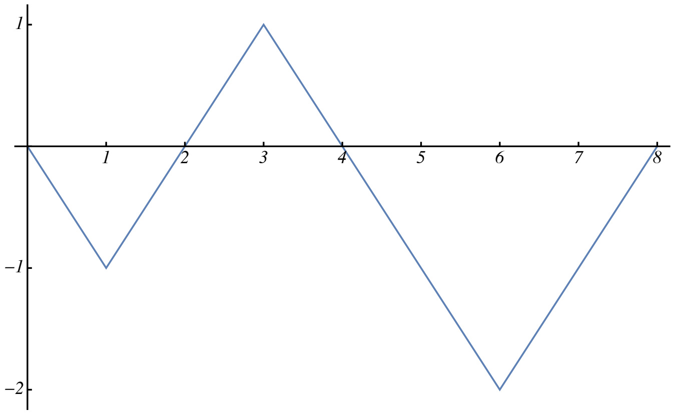

Figure 1. The first 8 steps of a random walk path. The horizontal axis marks time. The steps are $X_1 = -1, X_2 = 1, X_3 = 1, X_4 = -1, X_5 = -1, X_6 = -1, X_7 = 1, X_8 = 1$.

In random walk language, in the gambler's ruin problem we computed the probability that symmetric simple random walk started at x visits M before 0. We state this as a theorem. See Figure 2 for illustrations.

Theorem 3.60. *Fix integers $0 < x < M$. Consider SSRW $\{S_n\}_{n \ge 0}$ with nonrandom initial state $S_0 = x$. Then*

$$P(S_n \textit{ visits point } M \textit{ before visiting } 0) = \frac{x}{N}.$$

In the rest of the section we assume $S_0 = 0$. This is not a restriction if our starting point is deterministic: a random walk started from x looks the same as a random walk started from 0, and then shifted by x.

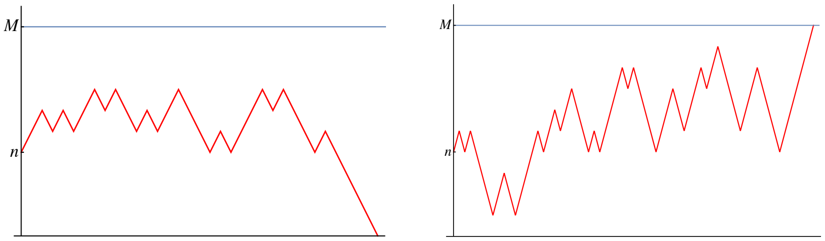

Figure 2. Two realizations of the random walk paths of the gambler's ruin problem. The first graph shows a losing outcome, while the second one shows a winning outcome.

A useful property of random walk is that, because future steps are independent of the past, the future of the random walk relative to its present position looks always the same as a fresh new random walk started at the origin.

Theorem 3.61. *For times* $0 \le m < n$ *and points* $a, b \in \mathbb{Z}$*, SRW with initial point* $S_0 = 0$ *satisfies*

$$P(S_{n+m} = a + b \,|\, S_m = a) = P(S_n = b).$$

Proof. The proof comes from the independence of $S_m = X_1 + \cdots + X_m$ and the random variables $X_{m+1}, X_{m+2}, X_{m+3}, \ldots$.

$$\begin{aligned} P(S_{n+m} = a + b \,|\, S_m = a) &= \frac{P(S_{n+m} = a+b, S_m = a)}{P(S_m = a)} = \frac{P(S_m = a, \sum_{i=m+1}^{n+m} X_i = b)}{P(S_m = a)} \\ &= \frac{P(S_m = a)P(\sum_{i=m+1}^{n+m} X_i = b)}{P(S_m = a)} \\ &= P(\sum_{i=m+1}^{n+m} X_i = b) = P(\sum_{i=1}^{n} X_i = b) \\ &= P(S_n = b). \end{aligned}$$

□

Symmetric simple random walk.

We compute the distributions of some quantities associated with symmetric SRW. So henceforth take $p = \frac{1}{2}$. We begin by finding the probability mass function of S_n. This is straightforward, except for a small wrinkle from *parity*: S_n must be an odd number at odd times n, and an even number at even times n. This is evident because taking one step from an odd number goes to an even number, and vice versa.

To find the possible values of S_n, first note that any possible value must be between $-n$ and n. Suppose point a can be reached with k $(+1)$-steps and $n - k$ (-1)-steps. Then $a = k - (n - k) = 2k - n$ from which $k = \frac{1}{2}(n + a)$ and $n - k = \frac{1}{2}(n - a)$. This tells us that a is a possible value of S_n if and only if $-n \le a \le n$ and $n - a$ is even (which also implies that $n + a$ is even because $n + a = (n - a) + 2a$). Since the number of $(+1)$-steps is $\text{Bin}(n, \frac{1}{2})$ distributed, we can summarize this discussion in the formula for $a \in \mathbb{Z}$ and $n \in \mathbb{Z}_{>0}$:

$$P(S_n = a) = \begin{cases} \dbinom{n}{\frac{n+a}{2}} 2^{-n}, & \text{if } -n \le a \le n \text{ and } n-a \text{ is even} \\ 0, & \text{otherwise.} \end{cases} \tag{3.44}$$

Next we turn to more challenging tasks: the running maximum, first return time to zero, and time spent on the positive half line. A priori these appear difficult tasks, but we find that we can make progress by counting paths.

Introduce the notation $N_{n,\,a\to b}$ for the number of paths of length n that start at a and end at b. Using the same ideas as for the probability mass function of the random walk we derive

$$N_{n,\,a\to b} = \begin{cases} \dbinom{n}{\frac{1}{2}(n+b-a)}, & \text{if } n-(b-a) \text{ is even} \\ 0, & \text{otherwise.} \end{cases}$$

The following combinatorial lemma is key to unlocking some of the secrets of SSRW.

Theorem 3.62 (Reflection principle)**.** *Let a, b be integers with $b > \max(0, a)$. The number of paths of length n that go from 0 to a and visit point b along the way is equal to $N_{n,\,0\to 2b-a}$. In particular, for SSRW,*

$$P(S_n = a,\, S_k = b \text{ for some } k = 0, \dots, n) = P(S_n = 2b - a). \tag{3.45}$$

Proof. The proof is basically Figure 3. The picture depicts a bijection between all paths from 0 to $2b - a$ (green path) and those paths from 0 to a that visit point b (green path from 0 to b followed by blue path to a). The point $2b - a$ is the reflection of a about b and lies above b. Thus every path from 0 to $2b-a$ must visit b.

The bijection of paths is defined by first following a path to its first visit to b, and then reflecting the remainder of the path across b. This mapping is its own inverse (reflecting twice brings you back where you started), and maps between the two types of paths.

The probability statement comes by multiplying the path counts by 2^{-n} which is the probability of each path. □

Distribution of the running maximum. The running maximum of the random walk is defined by $M_n = \max(0, S_1, \dots, S_n)$ for $n = 0, 1, 2, \dots$. It is always nonnegative. We find its distribution.

Theorem 3.63. *For $r \in \mathbb{Z}_{\ge 0}$, the running maximum of symmetric SRW satisfies*

$$P(M_n = r) = P(S_n = r) + P(S_n = r + 1).$$

Note that one of the terms on the right is always zero depending on the parity of $n - r$.

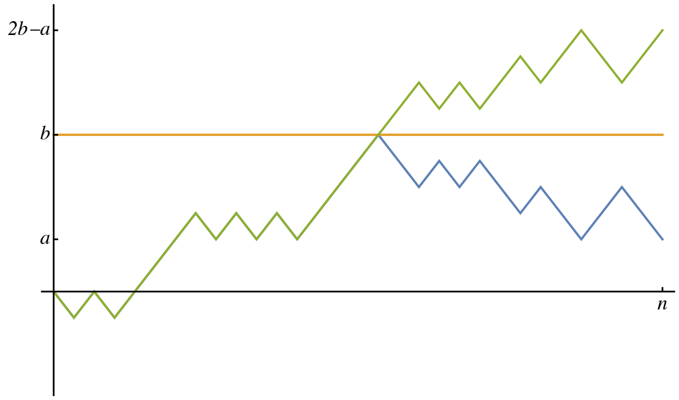

Figure 3. The figure shows the correspondence in the reflection principle. The green path connects 0 and $2b - a$ in n steps, crossing b (yellow horizontal line) along the way. The blue path is the reflection of the portion of the green path that starts from the first crossing of b. If we join the first portion of the green path with the blue path then we get a path connecting 0 and a in n steps that visits b along the way.

Proof. We begin by calculating the probabilities $P(M_n \geq r, S_n = a)$ for integers $r \geq 0$ and $a \in \mathbb{Z}$.

Suppose $a \geq r$. Then $S_n = a$ implies $M_n \geq r$, and so in this case

$$P(M_n \geq r, S_n = a) = P(S_n = a).$$

Suppose $a < r$. $M_n \geq r$ implies that the RW must have visited r. If $r > 0$ we can use the reflection principle as stated in Theorem 3.62 to get

$$\begin{aligned} P(M_n \geq r, S_n = a) &= P(S_n = a,\ S_k = r \text{ for some } k = 1, \dots, n-1) \\ &= P(S_n = 2r - a). \end{aligned}$$

The above identity holds also for $r = 0$ because we can drop the vacuous condition $M_n \geq 0$ and the identity above reduces to the true statement $P(S_n = a) = P(S_n = -a)$.

Since $\{M_n \geq r\}$ is the union of the disjoint events $\{M_n \geq r, S_n = a\}$ as a varies,

$$\begin{aligned} P(M_n \geq r) &= \sum_{a=-n}^{n} P(M_n \geq r, S_n = a) \\ &= \sum_{a=-n}^{r-1} P(M_n \geq r, S_n = a) + \sum_{a=r}^{n} P(M_n \geq r, S_n = a) \\ &= \sum_{a=-n}^{r-1} P(S_n = 2r - a) + \sum_{a=r}^{n} P(S_n = a) \\ &= P(S_n \geq r + 1) + P(S_n \geq r) \end{aligned}$$

Now we can get the probability mass function of the maximum:

$$\begin{aligned}
P(M_n = r) &= P(M_n \ge r) - P(M_n \ge r+1) \\
&= \big(P(S_n \ge r+1) + P(S_n \ge r)\big) - \big(P(S_n \ge r+2) + P(S_n \ge r+1)\big) \\
&= \big(P(S_n \ge r+1) - P(S_n \ge r+2)\big) + \big(P(S_n \ge r) - P(S_n \ge r+1)\big) \\
&= P(S_n = r+1) + P(S_n = r).
\end{aligned}$$

□

From Theorem 3.63 and equation (3.44) we derive a formula for the probability mass function of M_n:

$$P(M_n = r) = \begin{cases} \binom{n}{\frac{n+r}{2}} 2^{-n}, & \text{if } 0 \le r \le n \text{ and } n-r \text{ is even} \\ \binom{n}{\frac{n+r+1}{2}} 2^{-n}, & \text{if } 0 \le r \le n \text{ and } n-r \text{ is odd.} \end{cases} \tag{3.46}$$

Exercise 3.44 asks you to verify that this defines a legitimate probability mass function.

Returns to zero. From the probability mass function of S_n we get $P(S_n = 0) = \binom{n}{n/2}2^{-n}$ if n is even, $P(S_n = 0) = 0$ if n is odd. But what is the probability that the RW returns to zero the first time in the nth step?

We start by computing the probability that there are no visits to 0 within the steps $1, 2, \dots, 2k$. Surprisingly, this will be equal to the probability that we are at zero after the $2k$th step.

Theorem 3.64 (No returns to 0 within the first $2k$ steps)**.**

$$P(\textit{no visits to } 0 \textit{ within the first } 2k \textit{ steps}) = P(S_{2k} = 0) = \binom{2k}{k} 2^{-2k} \tag{3.47}$$

and

$$P(S_i \ne 0, i = 1, \dots, 2k) = \binom{2k}{k} 2^{-2k}. \tag{3.48}$$

Proof. If there are no visits to 0 up to $2k$ then the walk must stay on the positive or the negative side. By symmetry these have the same probability, so

$$P(S_i \ne 0, i = 1, \dots, 2k) = 2P(S_i < 0, i = 1, \dots, 2k).$$

If $S_i < 0$ for $i = 1, \dots, 2k$ then $S_1 = -1$. Hence

$$\begin{aligned}
P(S_i < 0, i = 1, \dots, 2k) &= P(S_i < 0, i = 2, \dots, 2k | S_1 = -1) P(S_1 = -1) \\
&= \frac{1}{2} P(S_i < 0, i = 2, \dots, 2k | S_1 = -1).
\end{aligned}$$

By shifting the starting point of path (imagine that you shift the origin of the coordinate system to $(1, -1)$), the conditional probability $P(S_i < 0, i = 2, \dots, 2k | S_1 = -1)$ can be rewritten as the unconditional probability $P(S_i \le 0, i = 1, 2, \dots 2k-1)$:

$$P(S_i < 0, i = 2, \dots, 2k | S_1 = -1) = P(S_i \le 0, i = 1, 2, \dots 2k - 1).$$

But

$$\begin{aligned}P(S_i \le 0, i = 1, 2, \dots 2k-1) &= P(M_{2k-1} = 0) = P(S_{2k-1} = 0) + P(S_{2k-1} = 1)\\ &= P(S_{2k-1} = 1) = \binom{2k-1}{k-1} 2^{-(2k-1)}.\end{aligned}$$

Collecting everything we get

$$P(S_i \ne 0, i = 1, \dots, 2k) = 2 \cdot \frac{1}{2} \cdot \binom{2k-1}{k-1} 2^{-(2k-1)} = \binom{2k-1}{k-1} 2^{-(2k-1)}.$$

We can also check that

$$\binom{2k-1}{k-1} 2^{-(2k-1)} = \frac{(2k-1)!}{k!(k-1)!} 2^{-(2k-1)} = \frac{2k(2k-1)!}{k!k \cdot (k-1)!} 2^{-2k} = \binom{2k}{k} 2^{-2k},$$

which finishes the proof. □

Now we can compute the distribution of the first return.

Theorem 3.65. *For $k \ge 1$ we have*

$$P(\textit{the first return to zero happened in the } 2k\textit{th step}) = \frac{1}{2k-1}\binom{2k}{k}\frac{1}{2^{2k}}.$$

Proof. The event {the first return to zero happened in the $2k$th step} is the set difference of the events $A_k =$ {no visits to 0 within the first $2k$ steps} and $A_{k-1} =$ {no visits to 0 within the first $2k-2$ steps}, and we also have $A_k \subset A_{k-1}$. Thus

$$P(\text{the first return to zero happened in the } 2k\text{th step}) = P(A_{k-1}) - P(A_k).$$

By Fact 3.64 we have

$$P(A_{k-1}) - P(A_k) = P(S_{2k-2} = 0) - P(S_{2k} = 0) = \binom{2k-2}{k-1} 2^{-2(k-1)} - \binom{2k}{k} 2^{-2k}.$$

From the definition we can check that $\binom{2k-2}{k-1} = \frac{k}{2(2k-1)}\binom{2k}{k}$ and hence

$$\begin{aligned}\binom{2k-2}{k-1} 2^{-2(k-1)} - \binom{2k}{k} 2^{-2k} &= 4 \cdot \frac{k}{2(2k-1)}\binom{2k}{k} 2^{-2k} - \binom{2k}{k} 2^{-2k}\\ &= \frac{1}{2k-1}\binom{2k}{k}\frac{1}{2^{2k}},\end{aligned}$$

which completes the proof. □

Let us denote the first return to zero by T. Fact 3.65 shows that $P(T = 2k) = \frac{1}{2k-1}\binom{2k}{k}\frac{1}{2^{2k}}$.

Theorem 3.66.

$$P(S_{2k} = 0, S_i \ge 0, i = 1, \dots, 2k-1) = \frac{1}{k+1}\binom{2k}{k} 2^{-2k}.$$

Proof. Let C_k be the number of paths of length $2k$ that stay non-negative for the whole path, and return to 0 in the end. Then

$$P(S_{2k} = 0, S_i \ge 0, i = 1, \dots, 2k-1) = C_k 2^{-2k}.$$

Imagine that we add an up-step at the beginning of such a path, and a down-step at the end of the path. This way we got a path of length $2k + 2$ that returns to 0 at the end, and stays positive for steps $1, 2, \dots, 2k + 1$. This is a one-to-one correspondence: the number of such paths is also C_k. Note that such a path makes its first return to zero at $2k + 2$.

The probability of the first return to 0 happening at $2k + 2$ was computed in Theorem 3.65:

$$P(\text{first return to zero is at } 2k + 2) = \frac{1}{2k+1}\binom{2k+2}{k+1}2^{-(2k+2)}.$$

Hence the number of paths of length $2k+2$ that return to 0 at $2k+2$ the first time is $\frac{1}{2k+1}\binom{2k+2}{k+1}$. Exactly half of these paths stay positive for the steps $1, 2, \dots, 2k-1$ (the others stay negative), which means that

$$C_k = \frac{1}{2}\cdot\frac{1}{2k+1}\binom{2k+2}{k+1} = \frac{1}{k+1}\binom{2k}{k}.$$

Hence

$$P(S_{2k} = 0, S_i \geq 0, i = 1, \dots, 2k-1) = \frac{1}{k+1}\binom{2k}{k}2^{-2k}.$$

□

The number $C_k = \frac{1}{k+1}\binom{2k}{k}$ is called the kth Catalan number. It shows up in a number of combinatorics applications.

Example 3.67 (Ballot problem)**.** Suppose that $a > b > 0$. Evaluate the conditional probability $P(S_i > 0 \text{ for } 1 \leq i \leq a+b \mid S_{a+b} = a-b)$.

This problem is called the ballot problem. Imagine that there is an election between two candidates (A and B) with $a+b$ voters. Assume that the voters decide their votes by flipping a fair coin independently of each other and giving their votes one by one. Given that candidate A wins the election by receiving a votes, what is the conditional probability that she was leading throughout the election process? By considering votes to A up steps and votes to B down steps in a random walk, you can check that this is exactly the same problem.

We first compute $P(S_{a+b} = a-b, S_i > 0 \text{ if } 1 \leq i \leq a+b)$. Removing the first step (which has to be an up step), and shifting the path gives

$$\begin{aligned}
&P(S_{a+b} = a-b, S_i > 0 \text{ if } 1 \leq i \leq a+b)\\
&\qquad = \frac{1}{2}P(S_{a+b-1} = a-b-1,\ \text{no visits to } -1)\\
&= \frac{1}{2}\left(P(S_{a+b-1} = a-b-1) - P(S_{a+b-1} = a-b+1)\right)\\
&\qquad = \left(\binom{a+b-1}{b} - \binom{a+b-1}{b-1}\right)2^{-(a+b)}\\
&\qquad = 2^{-(a+b)}\binom{a+b}{b}\left(\frac{a}{a+b} - \frac{b}{a+b}\right)\\
&= \frac{a-b}{a+b}P(S_{a+b} = a-b).
\end{aligned}$$

From this it follows that $P(S_i > 0 \text{ if } 1 \leq i \leq a+b | S_{a+b} = a-b) = \frac{a-b}{a+b}$. △

Time spent at the positive half line. We say that the kth step of the random walk is on the positive (negative) side, if both S_k and S_{k+1} are nonnegative (nonpositive). We define the time spent on the positive side within the first m steps by the number of steps on the positive side and denote it by I_m. Our goal is to determine the distribution of this random variable when $m = 2n$ is even.

Theorem 3.68 (Time spent on the positive side)**.** *Let $0 \le k \le n$. Then*

$$P(I_{2n} = 2k) = P(S_{2k} = 0)P(S_{2n-2k} = 0). \tag{3.49}$$

Proof. We have

$$P(I_{2n} = 2n) = P(S_i \ge 0, i = 1, \dots, 2n) = P(S_{2n} = 0)$$

by (3.48) of Fact 3.64.

By symmetry, $P(I_{2n} = 2n) = P(S_{2n} = 0)$. For $0 < k < n$ if we have $I_{2n} = 2k$ then the first return to zero is at most $2n - 2$ and even. Denoting this first return by T we have

$$P(I_{2n} = 2k) = \sum_{a=1}^{n-1} P(T = 2a)P(I_{2n} = 2k|T = 2a) \tag{3.50}$$

Until the first return the random walk fully stays on the positive side or the negative side, depending on the sign of the first step.

If $S_1 = -1$ then the random walk stays on the negative side for the first $2a$ steps, and all $2k$ steps after the $(2a)$th step. Since we are at 0 at the $(2a)$th step, the rest of the path is just a random walk of length $2n - 2a$ started at 0, hence $P(I_{2n} = 2k|T = 2a, S_1 = -1) = P(I_{2n-2a} = 2k)$. Similarly, $P(I_{2n} = 2k|T = 2a, S_1 = 1) = P(I_{2n-2a} = 2k - 2a)$ (in this case we only have to take $2n - 2a$ steps on the positive side after the $(2a)$th step.Thus

$$P(I_{2n} = 2k|T = 2a) = \frac{1}{2}P(I_{2n} = 2k|T = 2a, S_1 = -1) + \frac{1}{2}P(I_{2n} = 2k|T = 2a, S_1 = 1)$$

$$= \frac{1}{2}P(I_{2n-2a} = 2k) + \frac{1}{2}P(I_{2n-2a} = 2k - 2a).$$

Plugging this into (3.50) gives

(3.51)
$$P(I_{2n} = 2k) = \sum_{a=1}^{n-1} P(T = 2a)\left(\frac{1}{2}P(I_{2n-2a} = 2k) + \frac{1}{2}P(I_{2n-2a} = 2k - 2a)\right).$$

We will prove (3.49) by induction on k. As we have seen above, the statement is true for $k = 0$ for all n. Suppose that we have proved the statement for all values less than k (for all n), we will show that it must hold for k as well.

Suppose that by induction we have already shown the formula for values less than k. Then from (3.51) we get

$$\begin{aligned}P(I_{2n}=k) &= \frac{1}{2}\sum_{a=1}^{n-1} P(T=2a)\left(P(S_{2k}=0)P(S_{2n-2a-2k}=0)+P(S_{2k-2a}=0)P(S_{2n-2k}=0)\right)\\ &= \frac{1}{2}P(S_{2k}=0)\sum_{a=1}^{n-1}P(T=2a)P(S_{2n-2a-2k}=0)\\ &\quad +\frac{1}{2}P(S_{2n-2k}=0)\sum_{a=1}^{n-1}P(T=2a)P(S_{2k-2a}=0)\end{aligned}$$

Conditioning on the first return to 0 (as in the proof of (3.50)) gives

$$P(S_{2b}=0)=\sum_{i=1}^{b}P(T=2i)P(S_{2b-2i}=0).$$

and plugging this into the previous equation leads to

$$\begin{aligned}P(I_{2n}=k) &= \frac{1}{2}P(S_{2k}=0)P(S_{2n-2k}=0)+\frac{1}{2}P(S_{2k}=0)P(S_{2n-2k}=0)\\ &= P(S_{2k}=0)P(S_{2n-2k}=0)\end{aligned}$$

as we claimed. This proves the induction step and hence finishes the proof. □

3.6. Further mathematical issues ♣

Convolution.

In Section 3.3 we took the well-definedness of the convolutions for granted. In the discrete case this is straightforward. If p_X and p_Y are probability mass functions, then so is $p_X * p_Y$:

$$\begin{aligned}\sum_{n\in\mathbb{Z}} p_X*p_Y(n) &= \sum_{n\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} p_X(k)p_Y(n-k) = \sum_{k\in\mathbb{Z}} p_X(k)\sum_{n\in\mathbb{Z}} p_Y(n-k)\\ &= \sum_{k\in\mathbb{Z}} p_X(k)\sum_{m\in\mathbb{Z}} p_Y(m) = 1\cdot 1 = 1.\end{aligned}$$

In particular, it follows that all values $p_X * p_Y(n)$ are nonnegative finite numbers.

The situation is more complicated for functions. The value $f_X * f_Y(z)$ of a convolution of two probability density functions can actually blow up to infinity. (An example is in Exercise 3.52.) However, despite some blow-ups, we can check that $f_X * f_Y$ is a legitimate probability density function:

$$\begin{aligned}&\int_{-\infty}^{\infty} f_X*f_Y(z)\,dz = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\,dx\,dz\\ &= \int_{-\infty}^{\infty} f_X(x)\Big(\int_{-\infty}^{\infty} f_Y(z-x)\,dz\Big)\,dx = \int_{-\infty}^{\infty} f_X(x)\Big(\int_{-\infty}^{\infty} f_Y(y)\,dy\Big)\,dx\\ &= \Big(\int_{-\infty}^{\infty} f_X(x)\,dx\Big)\cdot\Big(\int_{-\infty}^{\infty} f_Y(y)\,dy\Big) = 1.\end{aligned}$$

Above we changed order of integration. This is always allowed for nonnegative functions by the Fubini-Tonelli theorem [Fol99, Theorem (2.37)]. In general, Young's inequality [Fol99, Theorem (8.7)] shows that the convolution of two absolutely integrable functions is again an absolutely integrable function.

Constructing independent random variables with specific distributions. Suppose that we have random variables $X_1, X_2, \dots, X_n$ that are defined on possibly different probability spaces $(\Omega_i, \mathcal{F}_i, P_i)$, $1 \le i \le n$. (That means that $X_i : \Omega_i \to \mathbb{R}$.) We would like realize these random variables in the same probability space, in a way that they become independent. More precisely, we want to define a probability space $(\Omega, \mathcal{F}, P)$ and random variables $Y_i : \Omega \to \mathbb{R}$, $1 \le i \le n$ so that Y_i has the same distribution as X_i, and $Y_1, \dots, Y_n$ are independent.

Here is one way to do this. Let the sample space be the direct product of the Ω_i:

$$\Omega = \Omega_1 \times \cdots \times \Omega_n = \{(\omega_1, \dots, \omega_n) : \omega_i \in \Omega_i\}.$$

To define the events we follow our usual strategy: we identify 'nice' events. In this case the 'nice' events are the subsets of Ω that can be written as direct product of events from $\mathcal{F}_i$: set of the form $A_1 \times \cdots \times A_n$ with $A_i \in \mathcal{F}_i$. These will all be events in our new probability space, and the actual $\mathcal{F}$ is the collection of sets that we get by including all other sets that we can construct from these using the defining properties of the σ-field.

We define P on the product sets the following way:

$$P(A_1 \times \cdots \times A_n) = P_1(A_1) P_2(A_2) \cdots P_n(A_n)$$

where we use the probability measures from the original probability spaces $(\Omega_i, \mathcal{F}_i, P_i)$ on the right. It is a non-trivial result of measure theory that one can extend the definition of P to all events in $\mathcal{F}$ so that it satisfies the requirements of a probability measure. The resulting probability measure is called the product-measure. (This is what we used to construct the probability space for infinitely many fair die rolls in Example 1.12, except there we used the direct product of infinitely many probability spaces.)

We now define the random variables $Y_i : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F}, P)$ as follows:

$$Y_i(\omega_1, \dots, \omega_n) = X_i(\omega_i).$$

To see that Y_i has the same distribution as X_i note that

$$\begin{aligned} P(Y_i \in B) &= P(\{(\omega_1, \dots, \omega_n) : X_i(\omega_i) \in B\}) \\ &= P(\Omega_1 \times \cdots \times \Omega_{i-1} \times X_i^{-1}(B) \times \Omega_{i+1} \times \cdots \times \Omega_n) \\ &= P(X_i^{-1}(B)) = P_i(X_i \in B), \end{aligned} \tag{3.52}$$

where $X_i^{-1}(B) = \{\omega_i \in \Omega_i : X(\omega_i) \in B\}$ is the inverse image of B with respect to X_i.

For the independence of $Y_1, \dots, Y_n$ we use the definition:

$$\begin{aligned} P(Y_1 \in B_1, \dots, Y_n \in B_n) &= P(X_1^{-1}(B_1) \times \cdots \times X_n^{-1}(B_n)) \\ &= P_1(X_1^{-1}(B_1)) \cdots P_n(X_n^{-1}(B_n)) \\ &= P_1(X_1 \in B_1) \cdot P_n(X_n \in B_n) \\ &= P(Y_1 \in B_1) \cdots P(Y_n \in B_n) \end{aligned}$$

where we used (3.52).

The same construction works for infinitely many probability spaces and random variables as well. In that case $\Omega = \Omega_1 \times \Omega_2 \times \cdots$ is the infinite product space, the 'nice' events are of the form $A_1 \times \cdots \times A_n \times \Omega_{n+1} \times \Omega_{n+2} \times \cdots$, with $A_i \in \Omega_i, 1 \le i \le n$, but otherwise the construction is the same.

Exercises

Exercise 3.1. Let $m \ge 3$. An urn contains m balls labeled $1, \dots, m$. Draw all the balls from the urn one by one without replacement and observe the labels in the order in which they are drawn. Let X_j be the label of the jth draw, $1 \le j \le m$. Assume that all orderings of the m draws are equally likely. Fix two distinct labels $a, b \in \{1, \dots, m\}$. Let

$$N = \min\{n \in \{1, \dots, m\} : X_n \in \{a, b\}\}$$

be the index of the first draw that is either a or b. Let $Y = X_N \in \{a, b\}$ be the label of this draw. Find the joint probability mass function of (N, Y) and the marginal probability mass functions of N and Y. Are N and Y independent random variables?

Exercise 3.2. Suppose that X_1, X_2, X_3 are independent random variables (see Def. 3.1). Prove the following direction of Theorem 3.4: for all $x_1, x_2, x_3 \in \mathbb{R}$ we have

$$F_{X_1,X_2,X_3}(x_1, x_2, x_3) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot F_{X_3}(x_3).$$

Exercise 3.3. Prove Theorem 3.19.
Hint. Adapt the proof described in Example 1.22: show that with probability one there is an $\ell \ge 0$ so that $X_{\ell k+1} = t_1, X_{\ell k+2} = t_2, \dots X_{\ell k+k} = t_k$.

Exercise 3.4. Prove the following generalization of Theorem 3.19. Let S be a countable set of real numbers. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables so that the set of possible values of X_1 is S. Let $(t_1, \dots, t_k)$ be a finite sequence with $t_i \in S, 1 \le i \le k$. Show that with probability one there is an $\ell \ge 0$ so that $X_{\ell k+1} = t_1, X_{\ell k+2} = t_2, \dots X_{\ell k+k} = t_k$.

Exercise 3.5. Prove Theorem 3.6 for discrete random variables. Assume that $p_1, p_2, \dots, p_n$ are nonnegative functions on $\mathbb{R}$. Let $X_1, X_2, \dots, X_n$ be discrete random variables defined on the same $(\Omega, \mathcal{F}, P)$. Assume that

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = p_1(x_1)p_2(x_2) \cdots p_n(x_n)$$

for all choices of x_i, for $i = 1, \dots, n$. Show that $X_1, \dots, X_n$ are independent, and there are positive constants $c_k, 1 \le k \le n$ so that $\prod_{j=1}^n c_j = 1$ and $p_k(x) = c_k P(X_k = x)$ for all x.
Hint. Adapt the proof for the $n = 2$ case.

Exercise 3.6. Suppose that X, Y are jointly continuous with joint probability density function

$$f(x, y) = \begin{cases} xe^{-x(1+y)}, & \text{if } x > 0 \text{ and } y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Determine whether X and Y are independent or not.

Exercise 3.7. Suppose a circular dart board is 18 inches in diameter, and the bull's eye in the center is 2 inches in diameter. What is the probability that out of four darts thrown at the dart board, exactly two hit the bull's eye? Assume that each dart lands at a uniformly random point on the board, independently of the other darts.

Exercise 3.8. Show that in an infinite sequence of fair die rolls, we eventually see two sixes in a row.

Hint. Imitate the beginning of the proof of Theorem 3.19, look at distinct 2-blocks of die rolls, and count arrangements.

Exercise 3.9. Imitate the proof of given for a case of Theorem 3.19 to show that in an infinite sequence of fair die rolls, we eventually see the numbers $1, 2, 3, 4, 5, 6$ appear in order one immediately after the other. To create a readable proof, introduce vector notation for both random variables and outcomes.

Exercise 3.10. It is important to understand that once independence of trials is given up, the almost sure (that is, with probability one) appearance of every outcome can be lost, even if there are infinitely many trials and even if marginally the probability of success does not change. Here are two examples.

(a) A magical coins behaves as follows. The first flip is fair, but after that every future flip repeats the first flip. In mathematical terms, let A_i be the event that the ith flip is tails, and assume that

$$P(A_1) = P(A_1^c) = \tfrac{1}{2} \quad \text{and} \quad P(A_i \,|\, A_1) = P(A_i^c \,|\, A_1^c) = 1 \quad \text{for } i \geq 2.$$

Show that $P(A_i) = \frac{1}{2}$ for all $i \geq 1$. What is the probability that tails is seen at least once after the mth flip, if the coin is flipped forever?

(b) 90% of the coins in circulation are fair, and the remaining 10% are coins so biased that *every* flip is tails. One coin is chosen randomly, and then this same coin is flipped forever. What is the probability of heads on the ith flip? What is the probability that heads is seen at least once after the mth flip?

For Exercises 3.11–3.13, consider a sequence of independent trials with success probability $p \in (0, 1)$, with random variables X_k defined as in (3.14).

Exercise 3.11. Let K be the index of the first successful trial that is immediately followed by a failure. In symbols,

$$K = \inf\{n \in \mathbb{Z}_{>0} : X_n = 1, X_{n+1} = 0\}. \tag{3.53}$$

Find the probability mass function of K. Check that your answer is a legitimate probability mass function.

Hint. Decompose the event $\{K = m\}$ into disjoint components expressed in terms of the trial outcome variables $\{X_i\}$. Note that a success before m cannot be immediately followed by a failure.

Exercise 3.12. Let N be the index of the first success, as defined in (3.18). Let K be the index of the first success that is immediately followed by a failure, as defined in (3.53).

(a) Find the probability $P(X_{N+1} = 0,\, X_{N+2} = 1,\, X_{N+3} = 0)$. In English, the task is to find the probability that the three trials immediately following the first success yield a failure, a success, and a failure, in that order.

(b) Find the probability $P(X_{K+1} = 0,\, X_{K+2} = 1,\, X_{K+3} = 0)$. The task is to find the probability that the three trials immediately following the random index K yield a failure, a success, and a failure, in that order. Explain why your answer makes intuitive sense.

Hint. In each case, decompose the probability according to the value of N or K (whichever the case), express all events in terms of the random variables $\{X_i\}$ and use their independence.

Significance. It is tempting to solve part (a) simply by saying that after the first success the trials continue exacty as before, and so the answer is $(1-p)\cdot p\cdot(1-p)$. This is *true*. But it is an insight that must be supported by a rigorous proof, and that is the purpose of the exercise.

The point of part (b) is that the answer to (a) can fail if we change the definition of the random index a little. Hence the answer to part (a) is not universally true but depends on the properties of the random index. The key feature is that N *does not look into the future but* K *does.* This is so because we cannot know whether $K = m$ is true without knowing the next trial X_{m+1} (the future). Random indices or *random times* such as N that do not look into the future are called *stopping times*. They are studied in the theory of stochastic processes. The result of part (a) is a very special case of the *strong Markov property*.

Exercise 3.13. Assume that $0 < p < 1$.

(a) Let $S_n \sim \mathrm{Bin}(n,p)$ count the number of successes in the first n trials. Fix a positive integer k. Show that $\lim_{n\to\infty} P(S_n \le k) = 0$.

(b) Show that in infinitely many trials there are infinitely many successes with probability one.
Hint. The complementary event is

$$\{\text{only finitely many successes}\} = \bigcup_{k\in\mathbb{Z}_{>0}} \{\text{at most } k \text{ successes}\}.$$

Use part (a).

Exercise 3.14. Let X be a discrete random variable with possible values $\{0,1,2,\dots\}$ and the following probability mass function: $P(X=0) = \frac{4}{5}$ and for $k \in \{1,2,3,\dots\}$

$$P(X=k) = \tfrac{1}{10}\cdot\left(\tfrac{2}{3}\right)^k.$$

(a) Verify that the above is a probability mass function.

(b) For $k \in \{1,2,\dots\}$, find $P(X \ge k \mid X \ge 1)$. Conditional on $X \ge 1$, does X have a named distribution?

Exercise 3.15. A medical trial of 80 patients is testing a new drug to treat a certain condition. This drug is expected to be effective for each patient with probability p, independently of the other patients. You personally have two friends in this trial. Given that the trial is a success for 55 patients, what is the probability that it was successful for both of your two friends?

Exercise 3.16. Flip a fair coin. If it is heads, stop. If it is tails, roll a fair die once. Then repeat. Keep flipping the coin until the first heads, and after each coin flip that is tails, roll the die once. Let X be the total number of sixes rolled. Find the probability mass function of X.

Hint. Let N be the number of flips needed to see the first heads. Deduce the probabilities $P(X = k \,|\, N = n)$ from the description of the game. To calculate $P(X = k)$ explicitly, use the series expansion (C.8), or evaluate explicitly the generating function $f(t) = \sum_{k=0}^{\infty} P(X = k)t^k$ and find its coefficients.

Exercise 3.17. Let $0 < p < 1$. Show that

$$\sum_{n=k}^{\infty} \binom{n-1}{k-1} p^k (1-p)^{n-k} = 1$$

by appeal to the series expansion (C.8) with $\alpha = -k$.

Exercise 3.18. For each positive integer n the following experiment is performed: $3n$ points are chosen independently of each other and uniformly at random on the interval $[0, n]$.

(a) Let X_n be the number of these random points in the interval $[2, 4]$. For $n \geq 4$ identify the distribution of X_n by name. For each $k \in \mathbb{Z}_{\geq 0}$, find the limit of $P(X_n = k)$ as $n \to \infty$.

(b) Let Y_n be the location of the leftmost point, that is,

$$Y_n = \sup\{h \in [0, n] : \text{there are no points in } [0, h)\}.$$

Find the c.d.f. F_n and the p.d.f. f_n of Y_n. For each fixed x, find the limit $F(x) = \lim_{n\to\infty} F_n(x)$. Is the limit the c.d.f. of some probability distribution? If so, which one?

Exercise 3.19. Let $Y \sim \text{Poisson}(\lambda)$. Calculate the probability $P(Y \text{ is even})$. **Hint.** Consider the series expansions of e^{λ} and $e^{-\lambda}$.

Exercise 3.20. A soccer player scores at least one goal in roughly half of her games. How would you estimate the percentage of games where she scores exactly three goals?

Exercise 3.21. Let $p_1, \dots, p_r$ be positive rational numbers such that $p_1 + \cdots + p_r = 1$. Devise an urn sampling problem that produces i.i.d. random variables $\{Y_k\}_{k=1}^{n}$ whose marginal distributions are $P(Y_k = j) = p_j$ for $j = 1, \dots, r$.

Hint. Adapt Example 3.15.

Exercise 3.22. Let $0 < p < 1$ and $0 < r < 1$ with $p \neq r$. You repeat a trial with success probability p until you see the first success. I repeat a trial with success probability r until I see the first success. All the trials are independent of each other.

(a) What is the probability that you and I performed the same number of trials?

(b) Let Z be the total number of trials you and I performed altogether. Find the possible values and the probability mass function of Z.

Exercise 3.23. Let $X \sim \text{Negbin}(k, p)$ and $Y \sim \text{Geom}(p)$ be independent random variables. Find the distribution of $X + Y$.

Exercise 3.24. Let $X \sim \text{Negbin}(k, p)$ and $Y \sim \text{Negbin}(\ell, p)$ be independent random variables. Find the distribution of $X + Y$.

Exercise 3.25. Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ be independent random variables. Find the distribution of $X + Y$.
Hint. You can do this without any computation.

Exercise 3.26. Let X be a uniformly chosen integer from the set $\{1, 2, \dots, n\}$ and Y be an independent uniformly chosen integer from the set $\{1, 2, \dots, m\}$ where $n < m$. Find the probability mass function of $X + Y$.

Exercise 3.27. Let X have density $f_X(x) = 2x$ for $0 < x < 1$ and let Y be uniform on the interval $(1, 2)$. Assume X and Y independent.

(a) Give the joint density function of (X, Y). Calculate $P(Y - X \geq \frac{3}{2})$.

(b) Find the density function of $X + Y$.

Exercise 3.28. Let X be a uniform random variable on the interval $[0, 1]$ and Y a uniform random variable on the interval $[8, 10]$. Suppose that X and Y are independent. Find the density function f_{X+Y} of $X+Y$ and sketch its graph. Check that your answer is a legitimate probability density function.

Exercise 3.29. Let $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ be independent random variables with $\lambda \neq \mu$. Compute the probability density function of $X + Y$.

Exercise 3.30. Let $X_1, X_2, \dots, X_n$ be i.i.d. random variables with $X_i \sim \text{Exp}(\lambda)$. Find the probability density function of $X_1 + \cdots + X_n$.
Hint. Use induction on n.

Exercise 3.31. Let X and Y be independent exponential random variables with parameter 1.

(a) Calculate the probability $P(Y \geq X \geq 2)$.

(b) Find the density function of the random variable $X - Y$.

Exercise 3.32. Suppose that X and Y are independent positive random variables with probability density functions f_X and f_Y. Show that $Z = X/Y$ is a continuous random variable and find its probability density function.

Exercise 3.33. Let $X_1, \dots, X_{100}$ be independent absolutely continuous random variables that all have the same marginal density function. Find the probability that X_{20} is the 50th largest number among these 100 numbers.

Exercise 3.34. We have an urn with 20 red, 10 black and 15 green balls. We take a sample of 30, without replacement, one by one with order. Find the probability that the 3rd, 10th and 23rd picks are of different colors.

Exercise 3.35. Let $U_1, \dots, U_n$ be i.i.d. Unif$[0, 1]$ random variables. Let T_j be the 'rank' of U_j among the numbers $U_1, \dots, U_n$ (i.e. where U_j stands in the ordered version of these numbers), more precisely:

$$T_j = \#\{i : 1 \leq i \leq n, U_i \leq U_j\}.$$

Find the joint distribution of $T_1, \dots, T_n$.

Exercise 3.36. Let $X_1, \dots, X_{20}$ a sample without replacement from the set $\{1, \dots, 100\}$. Find the probability that $X_3 > X_{10}$.

Exercise 3.37. Describe all joint distributions of X_1, X_2 which are exchangeable and only take values from $\{0, 1\}$.

Exercise 3.38. Describe all joint distributions of X_1, X_2, X_3 which are exchangeable and only take values from $\{0, 1\}$.

Exercise 3.39. Let Θ be a discrete random variable taking values in $(0, 1)$. Let $U_1, \dots, U_n$ be i.i.d. random variables that are uniform on $[0, 1]$ and independent of Θ. Finally, let $Y_j = I(\Theta < U_j)$ for $1 \le j \le n$.
Show that $Y_1, \dots, Y_n$ are exchangeable.

Exercise 3.40. A fair die is rolled 10 times. Define indicator random variables

$$I_k = \begin{cases} 1, & \text{if rolls } k, k+1, k+2 \text{ yield } 1, 2, 3 \text{ in order} \\ 0, & \text{otherwise.} \end{cases}$$

Give an example of a subset of the random variables $I_1, I_2, \dots, I_8$ that is exchangeable. Give another example of a subset that is not exchangeable.

Exercise 3.41. Let $0 < p < 1$ and suppose that $X_1, X_2, \dots$ are i.i.d. Bernoulli random variables with success probability p. Let $a_1, a_2, \dots, a_k$ be elements of $\{0, 1\}$. Show that

$$P(\text{there is an } n \ge 1 \text{ so that } X_n = a_1, X_{n+1} = a_2, \dots, X_{n+k-1} = a_k) = 1.$$

This means that for any finite sequence of 0s and 1s (these are the a_is) with probability one we can find consecutive elements of the sequence $X_1, X_2, \dots$ that are equal to $a_1, \dots, a_k$, respectively. In particular, we will see k ones next to each other eventually with probability one.

Exercise 3.42 (Asymmetric gambler's ruin)**.** Fix an integer $M > 0$ and a parameter $p \in (0, 1)$ such that $p \neq \frac{1}{2}$. You play repeatedly a gamble where you win a dollar with probability p and lose a dollar with probability $1 - p$. Successive plays are independent. For integers $0 < x < M$, let r_x be the probability that you reach M before 0 if you have x dollars at the beginning. Calculate r_x for $0 < x < M$.

Exercise 3.43. Let $\{X_i\}$ be i.i.d. random variables with $P(X_i = -1) = P(X_i = 1) = \frac{1}{2}$. Let $S_n = 1 + X_1 + \dots + X_n$ be symmetric simple random walk with initial point $S_0 = 1$. Find the probability that S_n eventually hits the point 0.

Hint. Define the events $A = \{S_n = 0 \text{ for some } n\}$ and for $M > 1$, $A_M = \{S_n \text{ hits 0 before hitting } M\}$. Show that $A_M \nearrow A$.

Exercise 3.44. Verify that equation (3.46) defines a probability mass function.

Exercise 3.45. Give an alternate proof of Theorem 3.65 using the following outline. If the first step is 1, then the random walker has to stay on the positive side until the $(2k-1)$st step, has to return to 1 at that point, and the last step must be a down step. Thus

$$\begin{aligned} &P(\text{the first return to zero happened in the } 2k\text{th step}, S_1 = 1) \\ &\qquad = P(S_1 = 1, S_{2k-1} = 1, S_{2k} = 0, S_i > 0 \text{ if } 2 \le i \le 2k-2). \end{aligned}$$

Compute the probability on the right by the first and last steps, shifting the path and using the reflection lemma.

Exercise 3.46. Consider a (symmetric) random walk started at zero. Show that with probability one it will return to zero. In other words show that if T is the time of first return to zero after the first step then $P(T < \infty) = 1$. Here are two possible approaches.

(1) Fact 3.65 shows that $P(T = 2k) = \frac{1}{2k-1}\binom{2k}{k}\frac{1}{2^{2k}}$. Prove that $\sum_{k=1}^{\infty} P(T = 2k) = 1$. (This is doable if you write down the Taylor expansion of an appropriate function.)

(2) The random walk will be at 1 or -1 after its first step. Thus it is enough to show that a random walk started at 1 will always visit 0. (By symmetry the same is true if it starts from -1.) For $k > 0$ let A_k be the event that the walk started from 1 visits 0 before k. Show that $A_1, A_2, \ldots$ is an increasing sequence of events and the union is exactly the event that the walk visits 0 at some point. Use this with Theorem 1.25 to prove the statement.

Exercise 3.47. Find the probability that a random walk started from 0 visits b the first time in the nth step.

Exercise 3.48. Let Q_{2n} denote the time of last visit to 0 up to the first $2n$ steps. ($Q_{2n} = 0$ if the walk does not return to zero by the $2n$th step.) Find the probability mass function of Q_{2n}.

Exercise 3.49. Let S_n be the position of a symmetric random walk after n steps (with $S_0 = 0$). Denote by M_n the maximum of the walk within the first n steps: $M_n = \max(S_0, S_1, \ldots, S_n)$.

Compute the following probability:

$$P(S_{99} = -3, S_{200} = -2, M_{99} \le 1, M_{200} \le 2).$$

Hint: Draw a picture. Your answer will involve binomial coefficients which you do not need to simplify.

Exercise 3.50. Let S_n be the position of a symmetric random walker after n steps (with $S_0 = 0$). Find the probability that the random walk visits the position 10 first at the 100th step:

$$P(S_{100} = 10,\ \text{but } S_n \neq 10 \text{ for } 0 < n < 100) = ?$$

Exercise 3.51. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $X_i \sim \text{Unif}[-\frac{1}{2}, \frac{1}{2}]$. Find the probability density function of $X_1 + \cdots + X_n$.
Hint. Use induction on n.

Exercises for Section 3.6.

Exercise 3.52. Define this function on $\mathbb{R}$:

$$f(x) = \begin{cases} \frac{1}{8}|x|^{-3/4}, & 0 < |x| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Check that f is a probability density function. Calculate $(f * f)(0)$.

Chapter 4

Expectation

Section 4.1 takes as starting point formulas for expectations of discrete and absolutely continuous random variables. A mathematically rigorous general definition of the expectation comes in Section 4.3, where the special formulas of Section 4.1 are justified.

4.1. Calculating expectations

The calculation of the *expectation* $E(X)$ of a random variable X begins with two formulas for the basic types of random variables. For a discrete random variable

$$E(X) = \sum_k k\, P(X = k) \tag{4.1}$$

where the sum ranges over all the possible values k of X. For an absolutely continuous random variable X with density function f

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{4.2}$$

Alternative terms for the expectation are *mean* and *first moment*. The notation can be simplified to EX. Another conventional symbol for the expectation is the lower case Greek letter mu: $\mu = E(X)$.

Example 4.1. (Expectation of Bernoulli and indicator random variables.) Let $0 \le p \le 1$ and $X \sim \text{Ber}(p)$. Then

$$E[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = 1 \cdot p + 0 \cdot (1 - p) = p. \tag{4.3}$$

In equation (2.1) we defined an indicator function I_B of a set B. When B is an event on a sample space Ω, I_B is called the *indicator random variable* of B and defined for $\omega \in \Omega$ by

$$I_B(\omega) = \begin{cases} 1, & \omega \in B \\ 0, & \omega \notin B. \end{cases} \tag{4.4}$$

Since the event $\{I_B = 1\}$ is the same as the event B, $P(I_B = 1) = P(B)$ and I_B is a Bernoulli random variable with parameter $P(B)$. Its mean is then $E(I_B) = P(B)$. Thus every probability of an event is also the expectation of the corresponding indicator random variable. △

Example 4.2 (Mean of a uniform random variable)**.** Let $X \sim \text{Unif}[a, b]$. Then

$$E[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_a^b x\,\frac{1}{b-a}\,dx = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}.$$

△

Example 4.3 (Continuation of Example 2.26)**.** A dart lands uniformly at random on the disk of radius r_0. Let R denote the distance of the dart from the center of the disk. Find $E(R)$, the expected distance of the dart from the center of the disk.

The density function was found in Example 2.26:

$$f_R(s) = \begin{cases} 2s/r_0^2, & 0 \le s < r_0 \\ 0, & s < 0 \text{ or } s \ge r_0. \end{cases}$$

The expectation is

$$E(R) = \int_{-\infty}^{\infty} s\, f_R(s)\,ds = \int_0^{r_0} s \cdot \frac{2s}{r_0^2}\,ds = \tfrac{2}{3} r_0.$$

△

Formulas (4.1) and (4.2) are not satisfactory definitions as they stand, even for discrete and absolutely continuous random variables. Unless X is a discrete random variable with only finitely many possible values, any serious discussion of the infinite series on the right-hand side of (4.1) and of the integral on the right-hand side of (4.2) has to pay attention to their convergence. Rather than elaborate on these formulas, we present a summary of the general definition of the expectation and then state formulas (4.1) and (4.2) as theorems. A rigorous technical development of the definition and proofs of the claims made here are deferred to Section 4.3.

Summary of the definition of the expectation.

For the reader whose frame of reference is the Riemann integral from calculus, the definition of the expectation will appear very different. Rather than give a single formula for all cases as the Riemann integral does (see (C.14) in Appendix C), the definition of the expectation proceeds through three successively more general levels. Along the way, several auxiliary definitions are made.

Definition 4.4 (Definition of the expectation EX)**.** Consider random variables on some probability space $(\Omega, \mathcal{F}, P)$.

Step 1. A **simple random variable** is a discrete random variable with finitely many values. Let X be a nonnegative simple random variable. Then X is of the form

$$X(\omega) = \sum_{i=1}^{m} \alpha_i\, I_{A_i}(\omega) \tag{4.5}$$

where $\alpha_1, \dots, \alpha_m$ are its distinct nonnegative real values and the events $A_i = \{X = \alpha_i\}$ form a partition of Ω. Define the expectation of X by

$$EX = \sum_{i=1}^{m} \alpha_i P(A_i) = \sum_{i=1}^{m} \alpha_i P(X = \alpha_i). \tag{4.6}$$

Step 2. Let X be a $[0,\infty]$-valued random variable on Ω. That is, the values $X(\omega)$ are nonnegative reals and possibly ∞. Then the expectation of X is defined by

$$\begin{gathered} EX = \sup\{EY : Y \text{ is a simple random variable on } \Omega \text{ such that} \\ 0 \le Y(\omega) \le X(\omega) \text{ for all } \omega \in \Omega\}. \end{gathered}$$

This defines a value $EX \in [0,\infty]$.

Step 3. Let X be a $[-\infty,\infty]$-valued random variable on Ω. That is, the values $X(\omega)$ are real numbers, ∞ or $-\infty$. The random variable X is decomposed into its **positive part** X^+ and **negative part** X^- by defining $X^+(\omega) = X(\omega) \vee 0$ and $X^-(\omega) = (-X(\omega)) \vee 0$. (See this definition applied to real numbers in equation (C.1) in Appendix C.) Random variables X, $|X|$, X^+ and X^- satisfy the identities

$$X(\omega) = X^+(\omega) - X^-(\omega) \quad \text{and} \quad |X(\omega)| = X^+(\omega) + X^-(\omega). \tag{4.7}$$

The expectations $E(X^+)$ and $E(X^-)$ were defined in Step 2. If at least one of them is finite, the expectation of X is defined by

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

In this case we say that EX is well-defined. If both $E(X^+)$ and $E(X^-)$ are infinite, then EX is not defined. △

The algebraic point about extended reals in $[-\infty,\infty]$ is that, for a real $c \in \mathbb{R}$, operations $c + \infty = \infty$ and $c - \infty = -\infty$ are well-defined, but $\infty - \infty$ is undefined.

Next we take up the special cases of discrete and absolutely continuous random variables. The formulas in the two theorems that follow are *not* new definitions. They are consequences of Definition 4.4 and proved in Section 4.3.

Theorem 4.5. *Suppose X is a discrete random variable. Then the following expectations have well-defined values in $[0,\infty]$:*

$$E(X^+) = \sum_{k\ge 0} k\, P(X = k) \quad \textit{and} \quad E(X^-) = \sum_{k \le 0} (-k)\, P(X = k). \tag{4.8}$$

If at least one of these values is finite, then EX is well defined as the difference:

$$EX = \sum_{k\ge 0} k\, P(X = k) \; - \; \sum_{k\le 0} (-k)\, P(X = k). \tag{4.9}$$

The expectations in (4.8) are well-defined because a series with nonnegative terms either converges to a finite real number or diverges to ∞. In both cases the series has a value. The series formula $EX = \sum_k kP(X = k)$ in (4.1) can be viewed as a concise representation of (4.9) when this is well-defined.

Theorem 4.6. *Suppose X is an absolutely continuous random variable with density function f. Then the following expectations have well-defined values in $[0,\infty]$:*

$$E(X^+) = \int_0^\infty x\,f(x)\,dx \quad \textit{and} \quad E(X^-) = \int_{-\infty}^0 (-x)f(x)\,dx. \tag{4.10}$$

If at least one of these values is finite, then EX is well defined as the difference:

$$EX = \int_0^\infty x\,f(x)\,dx \;-\; \int_{-\infty}^0 (-x)f(x)\,dx). \tag{4.11}$$

The expectations in (4.10) are well-defined because an improper integral of a nonnegative function either converges or diverges to ∞. The integral formula $EX = \int_{-\infty}^{\infty} xf(x)\,dx$ in (4.2) summarizes (4.11) when EX is well-defined.

There are two situations where calculation of the expectation EX is always justified:

(i) X has a single sign: either $P(X \ge 0) = 1$ or $P(X \le 0) = 1$.

(ii) X is *absolutely integrable* which means that $E[\,|X|\,] < \infty$. Absolute integrability of X guarantees that both $E(X^+)$ and $E(X^-)$ are finite and hence $E(X)$ is a real number.

Example 4.7. Here is an example of an infinite expectation and an example of a nonexistent expectation.

(i) Suppose X has density function

$$f_X(x) = \begin{cases} \frac{1}{2}x^{-3/2}, & x \ge 1 \\ 0, & x < 1. \end{cases}$$

Then $X \ge 0$ so its expectation does exist and can be calculated by the improper integral:

$$\begin{aligned} EX &= \int_0^\infty xf_X(x)\,dx = \int_1^\infty \tfrac{1}{2}x^{-1/2}\,dx = \lim_{m\to\infty}\int_1^m \tfrac{1}{2}x^{-1/2}\,dx \\ &= \lim_{m\to\infty}(m^{1/2}-1) = \infty. \end{aligned}$$

(ii) Let Y have the symmetric density function

$$f_Y(x) = \begin{cases} \frac{1}{4}|x|^{-3/2}, & |x| \ge 1 \\ 0, & |x| < 1. \end{cases}$$

Then the same calculation as above shows that $E(Y^+) = E(Y^-) = \infty$: by (4.10),

$$E(Y^+) = \int_0^\infty xf_Y(x)\,dx = \int_1^\infty \tfrac{1}{4}x^{-1/2}\,dx = \infty$$

and on the negative side, with a change of variable $x = -y$,

$$E(Y^-) = \int_{-\infty}^0 (-y)f_Y(y)\,dy = \int_{-\infty}^{-1} \tfrac{1}{4}(-y)^{-1/2}\,dy = \int_1^\infty \tfrac{1}{4}x^{-1/2}\,dx = \infty$$

Thus EY does not exist. △

Properties of the expectation.

Next we present some properties of the expectation as theorems. Some of these properties can be verified for discrete random variables, or continuous random variables, or both, from formulas (4.1) and (4.2), and appear as exercises.

One of the most fundamental things to realize is that EX is entirely determined by the probability distribution of X. Random variables that are equal in distribution have the same expectation, regardless of what their probability spaces happen to look like. Recall Definition 2.48 from Section 2.3.

Theorem 4.8. *Let X and Y be two random variables equal in distribution. Then $EX = EY$, provided these expectations are well-defined.*

This theorem follows easily for discrete and absolutely continuous random variables from formulas (4.1) and (4.2) because equality in distribution implies that random variables have the same probability mass function or density function.

Corollary 4.9. *Suppose X and Y are defined on the same probability space $(\Omega, \mathcal{F}, P)$ and $P(X = Y) = 1$. Then $EX = EY$, provided these expectations are well-defined.*

The corollary follows from Theorem 4.8 because by Theorem 2.51 almost sure equality implies equality in distribution. Corollary 4.9 is useful because it allows us to ignore the values of a random variable on events of probability zero. The next discrete example illustrates.

Example 4.10. Suppose $\Omega = \{a, b, c\}$, $P\{a\} = P\{b\} = 1/2$ and $P\{c\} = 0$. Define the random variables X and Y on Ω by

$$X(a) = Y(a) = 1,\; X(b) = Y(b) = 3,\; X(c) = 7, \quad \text{and} \quad Y(c) = 9.$$

Then the expectations are

$$EX = \sum_k k\, P(X = k) = 1 \cdot \tfrac{1}{2} + 3 \cdot \tfrac{1}{2} + 7 \cdot 0 = 2$$

and

$$EY = \sum_k k\, P(Y = k) = 1 \cdot \tfrac{1}{2} + 3 \cdot \tfrac{1}{2} + 9 \cdot 0 = 2.$$

The point is that the values $X(c) = 7$ and $Y(c) = 9$ taken on the event $\{c\}$ of zero probability made no difference to the expectations. △

The expectation of a nonnegative random variable has a convenient alternative formula.

Theorem 4.11. *Suppose $P(X \geq 0) = 1$. Then*

$$E(X) = \int_0^\infty P(X > s)\, ds. \tag{4.12}$$

In the particular discrete case where $P(X \in \mathbb{Z}_{\geq 0}) = 1$, the formula can also be expressed as

$$E(X) = \sum_{k=0}^{\infty} P(X > k). \tag{4.13}$$

Formula (4.12) is valid for all nonnegative random variables. Exercise 4.5 asks for a verification of (4.13) from formula (4.1). We prove (4.12) for X that has a density function.

Proof of Theorem 4.11 in the absolutely continuous case. Suppose X has density function f and $P(X \ge 0) = 1$. Then

$$\begin{aligned}\int_0^\infty P(X > s)\,ds &= \int_0^\infty \Bigl(\int_s^\infty f(x)\,dx\Bigr)\,ds = \int_0^\infty \Bigl(\int_0^x ds\Bigr) f(x)\,dx \\ &= \int_0^\infty x f(x)\,dx = EX.\end{aligned}$$

Switching around the order of integration is always justified when the integrand is nonnegative. □

It is frequently necessary to calculate the expectation of a function of a random variable X or a function of a random vector $\mathbf{X}$. It is convenient that these can be calculated from the distributions of X and $\mathbf{X}$. We prove below both the discrete and the absolutely continuous case.

Theorem 4.12. *Let $\mathbf{X} = (X_1, \dots, X_d)$ be a random vector (or a scalar random variable, in case $d = 1$) and g a real-valued function defined on the range of $\mathbf{X}$. Assume that either $g \ge 0$ or that $g(\mathbf{X})$ is absolutely integrable, so that the expectation $E[g(\mathbf{X})]$ is well-defined. Then we have the following formulas for this expectation.*

(a) *Suppose $X_1, \dots, X_d$ are discrete random variables. Then*

$$E[g(\mathbf{X})] = \sum_{\mathbf{k}} g(\mathbf{k}) P(\mathbf{X} = \mathbf{k}) \tag{4.14}$$

where $\mathbf{k}$ ranges over the possible values of $\mathbf{X}$.

(b) *Suppose $X_1, \dots, X_d$ are jointly continuous random variables with joint density function f. Then*

$$E[g(\mathbf{X})] = \int_{\mathbb{R}^d} g(\mathbf{x}) f(\mathbf{x})\,d\mathbf{x}. \tag{4.15}$$

Proof. Part (a). The proof comes from (i) first decomposing the event $\{g(\mathbf{X}) = \mathbf{y}\}$ into the union of pairwise disjoint events $\{\mathbf{X} = \mathbf{x}\}$ over those $\mathbf{x}$ that satisfy $g(\mathbf{x}) = \mathbf{y}$ and then (ii) rearranging the sum. In the sums below $\mathbf{x}$ ranges over the possible values of $\mathbf{X}$ and $\mathbf{y}$ over the possible values of $g(\mathbf{X})$.

$$\begin{aligned}E[g(\mathbf{X})] &= \sum_{\mathbf{y}} \mathbf{y}\, P\{g(\mathbf{X}) = \mathbf{y}\} = \sum_{\mathbf{y}} \mathbf{y} \sum_{\mathbf{x}: g(\mathbf{x}) = \mathbf{y}} P(\mathbf{X} = \mathbf{x}) \\ &= \sum_{\mathbf{y}} \mathbf{y} \sum_{\mathbf{x}} I\{g(\mathbf{x}) = \mathbf{y}\} P(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}) \sum_{\mathbf{y}} \mathbf{y}\, I\{g(\mathbf{x}) = \mathbf{y}\} \\ &= \sum_{\mathbf{x}} P(\mathbf{X} = \mathbf{x}) g(\mathbf{x}).\end{aligned}$$

Part (b). We prove first the case $g \ge 0$. This time the proof uses formula (4.12) and then a switch in the order of integration. This is justified because the integrand is nonnegative.

$$\begin{aligned}E[g(\mathbf{X})] &= \int_0^\infty P\{g(\mathbf{X}) > s\}\,ds = \int_0^\infty \biggl(\int_{g(\mathbf{x})>s} f(\mathbf{x})\,d\mathbf{x}\biggr)\,ds \\ &= \int_0^\infty \biggl(\int_{\mathbb{R}^d} I_{\{g(\mathbf{x})>s\}} f(\mathbf{x})\,d\mathbf{x}\biggr)\,ds = \int_{\mathbb{R}^d} \biggl(\int_0^\infty I_{\{g(\mathbf{x})>s\}}\,ds\biggr) f(\mathbf{x})\,d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \biggl(\int_0^{g(\mathbf{x})} ds\biggr) f(\mathbf{x})\,d\mathbf{x} = \int_{\mathbb{R}^d} g(\mathbf{x})\,f(\mathbf{x})\,d\mathbf{x}.\end{aligned}$$

To complete the proof we decompose g into the difference of its positive and negative parts: $g(\mathbf{x}) = g^+(\mathbf{x}) - g^-(\mathbf{x})$ where $g^+(\mathbf{x}) = g(\mathbf{x}) \vee 0$ and $g^-(\mathbf{x}) = (-g(\mathbf{x})) \vee 0$. Take apart the expectation, apply the part proved above separately to g^+ and g^-, and put the integral together again:

$$\begin{aligned}E[g(\mathbf{X})] &= E[g^+(\mathbf{X})] - E[g^-(\mathbf{X})] = \int_{\mathbb{R}^d} g^+(\mathbf{x})\,f(\mathbf{x})\,d\mathbf{x} - \int_{\mathbb{R}^d} g^-(\mathbf{x})\,f(\mathbf{x})\,d\mathbf{x} \\ &= \int_{\mathbb{R}^d} \bigl(g^+(\mathbf{x}) - g^-(\mathbf{x})\bigr)\,f(\mathbf{x})\,d\mathbf{x} = \int_{\mathbb{R}^d} g(\mathbf{x})\,f(\mathbf{x})\,d\mathbf{x}.\end{aligned}$$

□

Formula (4.15) can be used whenever a random variable can be written as a function of another random variable with a density function. The next example illustrates its application to a random variable that is neither discrete nor absolutely continuous.

Example 4.13. Let $0 < b < \infty$. Find EX for the random variable X with cumulative distribution function

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/(3b) & \text{if } 0 \le x < b, \\ 1 & \text{if } x \ge b. \end{cases}$$

X is neither discrete nor absolutely continuous and has a pointmass at b:

$$P(X = b) = F(b) - F(b-) = \tfrac{2}{3}.$$

To compute EX, we find a function g and a uniform random variable U such that $X \overset{d}{=} g(U)$ so that we can use $EX = E[g(U)]$. If we let $U \sim \text{Unif}[0, 3b]$ then F_X coincides with F_U on $[0, b)$, while X assigns the remaining $\frac{2}{3}$ probability to the point b. Hence we define

$$g(u) = \begin{cases} u, & 0 \le u < b, \\ b, & b \le u \le 3b. \end{cases}$$

We check that $g(U)$ has the same distribution as X. Both $g(U)$ and X lie in $[0, b]$ with probability one, and for $x < b$,

$$P\{g(U) \le x\} = P\{U \le x\} = x/(3b) = P(X \le x).$$

This is enough for concluding that $F_{g(U)} = F_X$. Hence their expectations agree, and we can calculate

$$EX = E[g(U)] = \frac{1}{3b}\int_0^{3b} g(u)\,du = \frac{1}{3b}\left(\int_0^b u\,du + \int_b^{3b} b\,du\right) = \tfrac{5}{6}b.$$

The outcome above can be seen as a combination of the absolutely continuous and discrete *parts* of X. Namely, X has a density function $1/(3b)$ on $[0, b)$ and probability $\frac{2}{3}$ at b. Together these give

$$EX = \int_0^b x\,\frac{1}{3b}\,dx + bP(X = b) = \tfrac{1}{6}b + \tfrac{2}{3}b = \tfrac{5}{6}b.$$

△

The next theorem lists some general properties of the expectation.

Theorem 4.14. *Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$. Assume their expectations are well-defined. Then their expectations have the following properties.*

(i) *Linearity: if EX and EY are finite, then for any real numbers a, b, c,*

$$E[aX + bY + c] = aE(X) + bE(Y) + c. \tag{4.16}$$

(ii) *Monotonicity: if $P(X \le Y) = 1$, then $EX \le EY$. If $P(X = Y) = 1$ then $EX = EY$.*

(iii) $|EX| \le E|X|$.

(iv) *Suppose $P(X \ge 0) = 1$ and $EX = 0$. Then $P(X = 0) = 1$.*

Point (iv) above may seem mysterious, but it turns out surprisingly useful. See also Exercise 4.8 for a quick proof of it.

Proof of Theorem 4.14 for the discrete case. Assume that X and Y are discrete with joint probability mass function p and marginal probability mass functions p_X and p_Y.

Part (i). Apply (4.14) to the function $g(x, y) = ax + by + c$, rearrange the sum, and use properties of probability mass functions:

$$\begin{aligned}E[aX + bY + c] &= \sum_{k,\ell}(ak + b\ell + c)p(k,\ell)\\ &= a\sum_k k\sum_\ell p(k,\ell) + b\sum_\ell \ell\sum_k p(k,\ell) + c\sum_{k,\ell} p(k,\ell)\\ &= a\sum_k k\,p_X(k) + b\sum_\ell \ell\,p_Y(\ell) + c = aE(X) + bE(Y) + c\end{aligned}$$

Since the sums are assumed to be convergent series, the manipulations above are justified.

Part (ii). We do the case where the expectations are finite. Suppose $P(X \le Y) = 1$ which is the same as $P(Y - X \ge 0) = 1$. Thus all the possible values of the discrete random variable $Y - X$ are nonnegative. Consequently the sum below has only nonnegative terms:

$$EY - EX = E[Y - X] = \sum_k k\, P(Y - X = k) \geq 0.$$

The second statement follows from the first since $P(X = Y) = 1$ implies both $P(X \leq Y) = 1$ and $P(X \geq Y) = 1$.

Part (iii). From the inequality $-|X| \leq X \leq |X|$, linearity (part (i)) and monotonicity (part (ii)) we deduce

$$-E|X| = E[-|X|\,] \leq EX \leq E|X|$$

which implies $|EX| \leq E|X|$.

Part (iv). $P(X \geq 0) = 1$ implies that all the possible values of X are nonnegative. Thus $\sum_k k\, P(X = k)$ is a sum or series of nonnegative terms whose value is zero. Then every term $k\, P(X = k)$ must be zero, which forces $P(X = k) = 0$ for every nonzero k. Thus $P(X = 0) = 1$. □

Special cases of expectations $E[g(X)]$ have names. For positive integers k, the *kth moment* of X is $E(X^k)$, the *kth absolute moment* of X is $E(|X|^k)$, and the *kth central moment* of X is $E[(X - \mu)^k]$ where $\mu = EX$. The first moment is the same as the mean, the second moment is also called the *mean square*, the first central moment is zero, and the second central moment is the *variance*, also denoted by σ^2 and $\mathrm{Var}(X)$.

An important property of moments is that they form a hierarchy. If $r < s$ are positive reals and $E[|X|^s]$ is finite, then also $E[|X|^r]$ is finite, as follows from the next theorem (Exercise 4.6 gives a hint for the proof).

Theorem 4.15. *Let X be a nonnegative random variable and $r \in (1, \infty)$. Assume that $E[X^r] < \infty$. Then $EX < \infty$.*

The variance is important enough to be highlighted as a formula. If X has finite mean $\mu = EX$, then its variance is

$$\mathrm{Var}(X) = \sigma^2 = E[(X - \mu)^2\,]. \tag{4.17}$$

Expansion of the square inside the brackets and linearity of expectation give an alternative formula for the variance:

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2 - 2\mu X + \mu^2\,] = E[X^2] - 2\mu EX + \mu^2 \\ &= E[X^2] - \mu^2. \end{aligned} \tag{4.18}$$

The derivation above gives us both a useful formula and a useful inequality:

$$\mathrm{Var}(X) \leq E[X^2]. \tag{4.19}$$

The square root of the variance is the *standard deviation*:

$$\mathrm{SD}(X) = \sigma = \left\{E[(X - \mu)^2\,]\right\}^{1/2}.$$

Example 4.16 (Variance of a Bernoulli and indicator random variable). Let $0 \le p \le 1$. Recall that $X \sim \text{Ber}(p)$ has probability mass function $P(X = 1) = p$ and $P(X = 0) = 1 - p$ and expectation $E(X) = p$. Hence its variance is

$$\begin{aligned}\text{Var}(X) &= E\big[(X-p)^2\big] = (1-p)^2 \cdot P(X=1) + (0-p)^2 \cdot P(X=0) \\ &= (1-p)^2 p + p^2 \cdot (1-p) = p(1-p).\end{aligned}$$

In particular, for an indicator random variable as defined by (4.4), we have $\text{Var}(I_A) = P(A)P(A^c)$. △

Example 4.17 (Mean and variance of a Poisson random variable). Let $0 < \lambda < \infty$ and $X \sim \text{Poisson}(\lambda)$. The probability mass function of X is $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \in \mathbb{Z}_{\ge 0}$.

$$E[X] = \sum_{k=0}^{\infty} k\, e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} = \lambda \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \lambda \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda.$$

To get $E(X^2)$, we calculate first $E[X(X-1)]$ because it comes easily.

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\, e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-2)!} = \lambda^2 \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda^2.$$

Hence,

$$\begin{aligned}\text{Var}(X) &= E[X^2] - (E[X])^2 = E[X(X-1)] + E[X] - (E[X])^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda.\end{aligned}$$

△

Example 4.18 (Variance of a uniform random variable). Let $X \sim \text{Unif}[a, b]$. In Example 4.2 we found $E[X] = \frac{a+b}{2}$. Another integration and some algebraic simplification gives

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{b^3 - a^3}{3(b-a)} = \frac{1}{3}(b^2 + ba + a^2).$$

Then by formula (4.18),

$$\text{Var}(X) = E(X^2) - (E[X])^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}.$$

△

We record two basic properties of the variance in the next theorem.

Theorem 4.19. *Let X be a random variable.*

(i) *Suppose X has finite variance and a, b are real numbers. Then* $\text{Var}(aX + b) = a^2\,\text{Var}(X)$.

(ii) $\text{Var}(X) = 0$ *if and only if there exists a real number c such that $P(X = c) = 1$. When this happens, c is the mean of X.*

Proof. Part (i) comes from a calculation that utilizes linearity of expectation:

$$\begin{aligned}\operatorname{Var}(aX+b) &= E\big[\big(aX+b-E(aX+b)\big)^2\,\big] = E\big[\big(aX+b-aE(X)-b\big)^2\,\big]\\ &= E\big[a^2(X-E(X))^2\big] = a^2E\big[(X-E(X))^2\big] = a^2\operatorname{Var}(X).\end{aligned}$$

Part (ii). If $P(X=c)=1$ then $EX=c$, $E(X^2)=c^2$ and

$$\operatorname{Var}(X) = E(X^2)-(EX)^2 = c^2-c^2 = 0.$$

Conversely, if $0=\operatorname{Var}(X)=E[(X-\mu)^2\,]$, then by Theorem 4.14 $P(X=\mu)=1$. □

Linearity of expectation.

In this section we take a closer look at applications of the linearity of the expectation.

Theorem 4.20. *Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space. Then*

$$E[X_1+X_2+\cdots+X_n] = E[X_1]+E[X_2]+\cdots+E[X_n], \tag{4.20}$$

provided all the expectations on both sides are finite.

In short, the expectation of a sum is *always* the sum of expectations. In particular, the random variables X_i in the theorem can themselves be complicated functions of other random variables. For example, the theorem contains as a special case a situation like this: let $\mathbf{X}_i$, $1\le i\le n$, be random vectors defined on $(\Omega,\mathcal{F},P)$ and for each index i let g_i be a real-valued function defined on the range of X_i. Then, as long as the expectations below are finite,

$$\begin{aligned}&E\big[g_1(\mathbf{X}_1)+g_2(\mathbf{X}_2)+\cdots+g_n(\mathbf{X}_n)\big]\\ &\qquad\qquad = E[g_1(\mathbf{X}_1)]+E[g_2(\mathbf{X}_2)]+\cdots+E[g_n(\mathbf{X}_n)].\end{aligned} \tag{4.21}$$

The caveat in the theorem about all expectations being finite is needed because there are examples where all the expectations are well-defined but the right-hand side of (4.20) is not defined. Here is one.

Example 4.21. Let X be any nonnegative random variable such that $E(X)=\infty$. Let $Y=5-X$. Then $X+Y=5$ and $E[X+Y]=5$, but $E[X]+E[Y]=\infty-\infty$ is not defined. △

Through linearity we can compute fairly easily some expectations that would be tedious or complicated to compute from formulas (4.1) or (4.2). The following examples illustrate the usefulness of linearity in situations where there is exchangeability.

Example 4.22. (Expected value of the binomial.) Let $X\sim\operatorname{Bin}(n,p)$. Since the expectation EX is the same for all $\operatorname{Bin}(n,p)$ random variables, we can choose a convenient probability space. So we imagine that X is the number of successes among n independent trials with success probability p. Then we can represent X as $X=X_1+X_2+\cdots+X_n$ where $X_1,X_2,\ldots,X_n$ are i.i.d. $\operatorname{Ber}(p)$ variables. By linearity of expectation and by Example 4.1 applied to each X_i,

$$E[X] = E[X_1]+E[X_2]+\cdots+E[X_n] = np.$$

△

Example 4.23. We deal five cards from a deck of 52 without replacement. Let X denote the number of aces among the chosen cards. Find the expected value of X.

The distribution of X does not depend on whether we choose cards in order or without order, so we can assume there is a first card, a second card and so on. Let $I_i, i = 1, \dots, 5$, be the indicator of the event that the ith card is an ace, that is,

$$I_i = \begin{cases} 1, & \text{if the } i\text{th card is an ace} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$X = I_1 + I_2 + \cdots + I_5$$

and by the linearity of expectation

$$E[X] = E[I_1] + E[I_2] + \cdots + E[I_5].$$

The random variables $I_1, \dots, I_5$ are exchangeable because we are sampling without replacement. Then $E[I_i] = E[I_1]$ for each i which implies

$$E[X] = 5E[I_1].$$

But $E[I_1] = P(\text{the first card is ace}) = \frac{4}{52} = \frac{1}{13}$. Thus $E[X] = \frac{5}{13}$.

Note that this expectation calculation would be exactly the same if the cards were drawn with replacement.

The random variable X can be rewritten as a sum of indicators in a different way which leads to an alternate method to arrive at the same answer. Let us label the four aces in the deck with the numbers $1, 2, 3, 4$ and denote by $J_i, i = 1, 2, 3, 4$, the indicator of the event that the ith ace is among the five chosen cards. Then $X = J_1 + J_2 + J_3 + J_4$ and $E[X] = E[J_1] + E[J_2] + E[J_3] + E[J_4]$. Again, by symmetry,

$$E[X] = 4E[J_1] = 4P(\text{the ace of spades is among the five chosen cards}).$$

We compute the last probability by imagining an unordered sample of five cards:

$$P(\text{the ace of spades is among the five}) = \frac{\binom{1}{1}\binom{51}{4}}{\binom{52}{5}} = \frac{\frac{51\cdot 50\cdot 49\cdot 48}{4!}}{\frac{52\cdot 51\cdot 50\cdot 49\cdot 48}{5!}} = \frac{5}{52}.$$

We obtain again $E[X] = 4 \cdot \frac{5}{52} = \frac{5}{13}$. △

Example 4.24. (Expected number of triangles in the random graph $G(n,p)$.) In general, a *graph* is a pair $G = (V, E)$ where V is a set of vertices and E is a set of edges between vertices. Three vertices form a *triangle* in a graph if there is an edge between each pair of vertices.

The *Erdős-Rényi random graph* $G(n,p)$ is defined by taking $\{1, 2, \dots, n\}$ as the set of vertices and then drawing an edge between each unordered pair i, j $(i \neq j)$ independently with probability p. Let X denote the number of triangles in $G(n,p)$. We calculate EX.

For each triple of distinct vertices i, j, k let $I_{i,j,k}$ be the indicator random variable whose value is one if and only if vertices i, j, k form a triangle. Equivalently, $I_{i,j,k} = 1$ if and only there is an edge between i and j, between i and k, and between

j and k. Thus $X = \sum_{i,j,k \text{ distinct}} I_{i,j,k}$. Since $I_{i,j,k} \sim \text{Ber}(p^3)$ and there are $\binom{n}{3}$ triples of distinct vertices,

$$EX = \sum_{i,j,k \text{ distinct}} E(I_{i,j,k}) = \binom{n}{3} p^3.$$

Here is a more mundane story that leads to the same mathematics. There are n guests at a party. Each pair of guests know each other with probability p, independently of the other guests. What is the expected number of groups of size three where each person knows the other two? △

Expectation and independence.

Our earlier discussion of independence saw that independence was associated with products of probabilities. This feature extends also to expectations.

Theorem 4.25. *Suppose $X_1, \dots, X_n$ are independent random variables. Then for all functions $g_1, \dots, g_n$ for which the expectations below are well-defined,*

$$E\left[\prod_{k=1}^{n} g_k(X_k)\right] = \prod_{k=1}^{n} E\big[g_k(X_k)\big]. \tag{4.22}$$

Proof. The theorem is true for all random variables. We give a proof for two absolutely continuous random variables. So suppose X and Y are independent with density functions f_X and f_Y. Then their joint density function is $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. We compute

$$\begin{aligned} E[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy = E[g(X)] \cdot E[h(Y)]. \end{aligned}$$

In the last step we were able to separate the double integral into a product of two single variable integrals which are exactly the expected values of $g(X)$ and $h(Y)$. □

We utilize Theorem 4.25 to give us a result for the variance of a sum of independent random variables. This result will be generalized in Theorem 4.33 and Corollary 4.34 in Section 4.2.

Theorem 4.26. *Assume the random variables $X_1, \dots, X_n$ are independent and have finite variances. Then*

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n). \tag{4.23}$$

Proof. Let $\mu_i = E(X_i)$. Use the definition of the variance, expand the square inside the expectation, separate the diagonal terms where $i = j$, and use linearity of expectation:

$$\begin{aligned}\operatorname{Var}\Bigl(\sum_{i=1}^n X_i\Bigr) &= E\Bigl[\Bigl(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\Bigr)^2\Bigr] = E\Bigl[\Bigl(\sum_{i=1}^n (X_i-\mu_i)\Bigr)^2\Bigr] \\ &= E\Bigl[\sum_{1\le i,j\le n} (X_i-\mu_i)(X_j-\mu_j)\Bigr] \\ &= \sum_{i=1}^n E\bigl[(X_i-\mu_i)^2\bigr] + \sum_{i\ne j} E\bigl[(X_i-\mu_i)(X_j-\mu_j)\bigr] \\ &= \sum_{i=1}^n \operatorname{Var}(X_i).\end{aligned}$$

The cross terms vanish above by independence: for $i \ne j$,

$$E[(X_i-\mu_i)(X_j-\mu_j)] = E[X_i-\mu_i]\cdot E[X_j-\mu_j] = 0.$$

□

Example 4.27 (Variance of the binomial)**.** As in Example 4.22, represent $X \sim \operatorname{Bin}(n,p)$ as $X = X_1+\dots+X_n$ for independent $X_i \sim \operatorname{Ber}(p)$. For each i, $\operatorname{Var}(X_i) = p(1-p)$ by Example 4.16. By Theorem 4.26

$$\operatorname{Var}(X) = \operatorname{Var}(X_1)+\dots+\operatorname{Var}(X_n) = np(1-p).$$

△

4.2. Covariance and correlation

Covariance and correlation quantify the strength and type of dependence between two random variables.

Definition 4.28. Let X and Y be random variables defined on the same sample space with expectations μ_X and μ_Y. The **covariance** of X and Y is

$$\operatorname{Cov}(X,Y) = E\bigl[(X-\mu_X)(Y-\mu_Y)\bigr] \tag{4.24}$$

if the expectation on the right is finite.

By expanding the product inside the expectation, we get an alternative formula for the covariance:

$$\begin{aligned}\operatorname{Cov}(X,Y) &= E\bigl[(X-\mu_X)(Y-\mu_Y)\bigr] = E\bigl[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y\bigr] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y.\end{aligned}$$

When $X = Y$, the covariance is the variance:

$$\operatorname{Cov}(X,X) = E\bigl[(X-\mu_X)^2\bigr] = \operatorname{Var}(X). \tag{4.25}$$

Formula (4.24) reveals that the sign of $\operatorname{Cov}(X,Y)$ carries meaning. Inside the expectation we have

$$\begin{aligned} & (X-\mu_X)(Y-\mu_Y) > 0 \quad \text{if } X-\mu_X \text{ and } Y-\mu_Y \text{ have the same sign} \\ \text{while} \quad & (X-\mu_X)(Y-\mu_Y) < 0 \quad \text{if } X-\mu_X \text{ and } Y-\mu_Y \text{ have the opposite sign.}\end{aligned}$$

Thus $\mathrm{Cov}(X, Y) > 0$ if on average X and Y tend to deviate together above or below their means, while $\mathrm{Cov}(X, Y) < 0$ if on average X and Y tend to deviate in opposite directions. Here is the commonly used terminology: X and Y are

- *positively correlated* if $\mathrm{Cov}(X, Y) > 0$,
- *negatively correlated* if $\mathrm{Cov}(X, Y) < 0$, and
- *uncorrelated* if $\mathrm{Cov}(X, Y) = 0$.

We illustrate these three possibilities with indicator random variables.

Example 4.29. We compute $\mathrm{Cov}(I_A, I_B)$ for the indicator random variables of two events A and B. Observe that $I_A I_B = I_{A\cap B}$ since $I_A I_B = 1$ exactly when both indicator random variables are equal to one, which happens precisely on $A \cap B$.

$$\begin{aligned}\mathrm{Cov}(I_A, I_B) &= E[I_A I_B] - E[I_A]E[I_B] = E[I_{A\cap B}] - E[I_A]E[I_B]\\ &= P(A\cap B) - P(A)P(B) = P(B)[P(A|B) - P(A)].\end{aligned}$$

For the last equality assume that $P(B) > 0$ so that $P(A|B)$ makes sense.

Thus indicator random variables I_A and I_B are positively correlated if $P(A|B) > P(A)$, in other words, if the occurrence of B increases the chances of A, and negatively correlated if the occurrence of B decreases the chances of A. Furthermore, $\mathrm{Cov}(I_A, I_B) = 0$ if and only if $P(A \cap B) = P(A)P(B)$, which is exactly the definition of independence of A and B. In other words, *in the case of indicator random variables independence is equivalent to being uncorrelated.* △

Remark 4.30. (Uncorrelated versus independent random variables.) If X and Y are independent, their covariance vanishes:

$$\mathrm{Cov}(X, Y) = E[XY] - \mu_X\mu_Y = E[X]E[Y] - \mu_X\mu_Y = 0.$$

Thus independent random variables are uncorrelated. The converse *does not hold in general.* That is, there are uncorrelated random variables that are not independent. For an example, let X be uniform on the set $\{-1, 0, 1\}$ and $Y = X^2$. We leave it as Exercise 4.13 to check that X and Y are uncorrelated but not independent. △

The next theorem collects properties of the covariance.

Theorem 4.31. *The following statements hold when the covariances are well-defined.*

(i) $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.

(ii) $\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$ *for real numbers* a, b.

(iii) *Covariace is bilinear: namely, for random variables* X_i *and* Y_j *and real numbers* a_i *and* b_j*,*

$$\mathrm{Cov}\Big(\sum_{i=1}^m a_i X_i\,,\,\sum_{j=1}^n b_j Y_j\Big) = \sum_{i=1}^m\sum_{j=1}^n a_i b_j\,\mathrm{Cov}(X_i, Y_j). \tag{4.26}$$

Proof. The symmetry in (i) is immediate from definition (4.24). For (ii), we can use the linearity of expectation to get

$$\begin{aligned}\mathrm{Cov}(aX + b, Y) &= E[(aX + b)Y] - E[aX + b]E[Y]\\ &= aE[XY] + bE[Y] - aE[X]E[Y] - bE[Y] = a(E[XY] - E[X]E[Y])\end{aligned}$$

which gives the statement we wanted to show.

For (iii) let us introduce the notations $\mu_{X_i} = E[X_i]$, $\mu_{Y_j} = E[Y_j]$ and note that

$$E\biggl[\sum_{i=1}^m a_i X_i\biggr] = \sum_{i=1}^m a_i \mu_{X_i} \quad \text{and} \quad E\biggl[\sum_{j=1}^n b_j Y_j\biggr] = \sum_{j=1}^n b_j \mu_{Y_j}.$$

Then we do some algebra inside the expectation, and use linearity of expectation:

$$\begin{aligned}
\mathrm{Cov}\biggl(\sum_{i=1}^m a_i X_i\,,\,\sum_{j=1}^n b_j Y_j\biggr) &= E\biggl[\biggl(\sum_{i=1}^m a_i X_i - \sum_{i=1}^m a_i\mu_{X_i}\biggr)\biggl(\sum_{j=1}^n b_j Y_j - \sum_{j=1}^n b_j\mu_{Y_j}\biggr)\biggr] \\
&= E\biggl[\biggl(\sum_{i=1}^m a_i(X_i - \mu_{X_i})\biggr)\biggl(\sum_{j=1}^n b_j(Y_j - \mu_{Y_j})\biggr)\biggr] \\
&= E\biggl[\sum_{i=1}^m\sum_{j=1}^n a_i b_j (X_i - \mu_{X_i})(Y_j - \mu_{Y_j})\biggr] \\
&= \sum_{i=1}^m\sum_{j=1}^n a_i b_j E[(X_i - \mu_{X_i})(Y_j - \mu_{Y_j})] \\
&= \sum_{i=1}^m\sum_{j=1}^n a_i b_j \,\mathrm{Cov}(X_i, Y_j).
\end{aligned}$$ □

Example 4.32. (Multinomial random variables.) We have a trial with r possible outcomes, labeled $1, \dots, r$. Outcome i appears with probability p_i, and $p_1 + \cdots + p_r = 1$. We perform n independent repetitions of this trial and denote by X_i the number of appearances of the ith outcome. We calculate $\mathrm{Cov}(X_i, X_j)$.

From Example 3.41 we know that the marginal distribution of X_i is $\mathrm{Bin}(n, p_i)$, and so

$$\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = np_i(1 - p_i).$$

Consider $i \neq j$. We decompose the variables into sums of indicators to take advantage of bilinearity. Define the indicator random variables

$$I_{k,i} = \begin{cases} 1, & \text{if trial } k \text{ gives outcome } i \\ 0, & \text{if trial } k \text{ gives an outcome other than } i. \end{cases}$$

Then $X_i = \sum_{k=1}^n I_{k,i}$ and $X_j = \sum_{k=1}^n I_{k,j}$. Bilinearity of covariance gives

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}\biggl(\sum_{k=1}^n I_{k,i}\,,\,\sum_{\ell=1}^n I_{\ell,j}\biggr) = \sum_{k=1}^n\sum_{\ell=1}^n \mathrm{Cov}(I_{k,i}, I_{\ell,j}). \tag{4.27}$$

If $k \neq \ell$, $I_{k,i}$ and $I_{\ell,j}$ are independent because they depend on distinct trials, and trials are independent by design. Hence in the last sum on line (4.27), only terms with $k = \ell$ are nonzero and

$$\mathrm{Cov}(X_i, X_j) = \sum_{k=1}^n \mathrm{Cov}(I_{k,i}, I_{k,j}).$$

Since $i \neq j$, we have $I_{k,i}I_{k,j} = 0$ because trial k cannot simultaneously yield both outcome i and outcome j. We deduce

$$\operatorname{Cov}(I_{k,i}, I_{k,j}) = E[I_{k,i}I_{k,j}] - E[I_{k,i}]E[I_{k,j}] = 0 - p_i p_j = -p_i p_j.$$

Return back to line (4.27) to conclude:

$$\operatorname{Cov}(X_i, X_j) = \sum_{k=1}^{n} \operatorname{Cov}(I_{k,i}, I_{k,j}) = -np_i p_j \qquad \text{for } i \neq j.$$

In particular we see that if $i \neq j$, then X_i and X_j are negatively correlated. This is natural since the more often outcome i appears, the fewer opportunities there are for outcome j.

Here is an alternative quick solution. By (4.28),

$$\operatorname{Var}(X_i + X_j) = \operatorname{Var}(X_i) + \operatorname{Var}(X_j) + 2\operatorname{Cov}(X_i, X_j).$$

Since $X_i + X_j$ is binomial with parameter $p_i + p_j$,

$$n(p_i + p_j)(1 - p_i - p_j) = np_i(1 - p_i) + np_j(1 - p_j) + 2\operatorname{Cov}(X_i, X_j).$$

This gives $\operatorname{Cov}(X_i, X_j) = -np_i p_j$. △

Variance of a sum. With the covariance, we can express the variance of a sum of random variables succinctly, without any assumptions of independence.

Theorem 4.33. *Let $X_1, \dots, X_n$ be random variables with finite variances and covariances. Then*

$$\operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j). \tag{4.28}$$

Proof. The proof uses properties of the covariance. Begin with (4.25). The second sum uses index j so that the indices of the two sums can be kept separate. Then use bilinearity and separate the diagonal terms ($i = j$).

$$\begin{aligned}\operatorname{Var}\Big(\sum_{i=1}^{n} X_i\Big) &= \operatorname{Cov}\Big(\sum_{i=1}^{n} X_i\,,\, \sum_{j=1}^{n} X_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} \operatorname{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \operatorname{Cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2\sum_{1 \le i < j \le n} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

The last step restricted the summation to $i < j$. Then each pair $i \neq j$ appears only once, hence the factor 2 in the front. □

In the case of two random variables equation (4.28) simplifies to

$$\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X, Y). \tag{4.29}$$

When the last sum in (4.28) vanishes, we get the following corollary that generalizes the earlier Theorem 4.26.

Corollary 4.34. *Let $X_1, \dots, X_n$ be uncorrelated random variables with finite variances. Then*

$$\operatorname{Var}(X_1 + X_2 + \cdots + X_n) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + \cdots + \operatorname{Var}(X_n).$$

Example 4.35. An urn has five balls, three red ones and two green ones. Draw three balls one by one. Let X denote the number of red balls in the sample. Compute $\operatorname{Var}(X)$ when the sampling is done (a) with replacement and (b) without replacement.

Start by introducing indicator random variables for the outcomes of the draws:

$$Y_k = \begin{cases} 1, & \text{if the } k\text{th draw is red} \\ 0, & \text{if the } k\text{th draw is green} \end{cases}$$

so that $X = \sum_{k=1}^{3} Y_k$. Y_1 is a Bernoulli random variable with success probability $\frac{3}{5}$. By exchangeability, so are Y_2 and Y_3. Thus $\operatorname{Var}(Y_k) = \frac{3}{5} \cdot \frac{2}{5} = \frac{6}{25}$.

(a) In sampling with replacement, the Y_ks are independent. Consequently $\operatorname{Var}(X) = \sum_{k=1}^{3} \operatorname{Var}(Y_k) = 3 \cdot \frac{6}{25} = \frac{18}{25}$.

(b) In sampling without replacement, the Y_ks are not independent. Calculate their covariance: for $i \neq j$, by exchangeability,

$$\begin{aligned} \operatorname{Cov}(Y_i, Y_j) &= \operatorname{Cov}(Y_1, Y_2) = E[Y_1 Y_2] - E[Y_1]E[Y_2] \\ &= P(\text{first two draws are red}) - \tfrac{3}{5} \cdot \tfrac{3}{5} = \tfrac{3 \cdot 2}{5 \cdot 4} - \tfrac{9}{25} = -\tfrac{6}{100}. \end{aligned}$$

The covariance is negative because drawing a red ball reduces the chances of further red draws. We apply (4.28) to calculate $\operatorname{Var}(X)$:

$$\operatorname{Var}(X) = \sum_{k=1}^{3} \operatorname{Var}(Y_k) + 2 \sum_{i<j} \operatorname{Cov}(Y_i, Y_j) = 3 \cdot \tfrac{6}{25} - 6 \cdot \tfrac{6}{100} = \tfrac{36}{100}.$$

The qualitative conclusion is that there is less variability in the case of sampling without replacement. If we draw all 5 balls without replacement, there cannot be any variability in the number of red balls. In this case the calculation gives zero variance, as it should:

$$\operatorname{Var}\Bigl(\sum_{k=1}^{5} Y_k\Bigr) = 5 \operatorname{Var}(Y_1) + 20 \operatorname{Cov}(Y_1, Y_2) = 5 \cdot \tfrac{6}{25} - 20 \cdot \tfrac{6}{100} = 0.$$

△

Correlation. By a suitable normalization of the covariance we get the correlation coefficient. Recall from Theorem 4.19 that $\operatorname{Var}(X) > 0$ unless X is *degenerate* which means that $P(X = a) = 1$ for some real a.

Definition 4.36. Let X and Y be random variables such that $\operatorname{Cov}(X, Y)$ is finite, $0 < \operatorname{Var}(X) < \infty$ and $0 < \operatorname{Var}(Y) < \infty$. The **correlation coefficient** of X and Y is defined by

$$\operatorname{Corr}(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}}.$$

△

Alternative notation for the correlation coefficient is $\rho(X,Y)$ or simply ρ. The correlation coefficient has this feature: for a nonzero real number a,

$$\operatorname{Corr}(aX,Y) = \frac{\operatorname{Cov}(aX,Y)}{\sqrt{\operatorname{Var}(aX)}\sqrt{\operatorname{Var}(Y)}} = \frac{a\operatorname{Cov}(X,Y)}{|a|\sqrt{\operatorname{Var}(X)}\sqrt{\operatorname{Var}(Y)}} = \frac{a}{|a|}\operatorname{Corr}(X,Y).$$

In other words, the magnitude of the correlation coefficient does not change under a scaling of the random variables, for example by changing units of measurement. Thus it is a more meaningful quantification of the dependence between X and Y than the covariance. The next theorem shows that the extreme values of the correlation coefficient have a very precise interpretation, namely linear (or more precisely, affine) dependence.

Theorem 4.37. *Assume that* $\operatorname{Cov}(X,Y)$ *is finite,* $0 < \operatorname{Var}(X) < \infty$ *and* $0 < \operatorname{Var}(Y) < \infty$*. The correlation coefficient has these properties.*

(i) $-1 \le \operatorname{Corr}(X,Y) \le 1$.

(ii) $\operatorname{Corr}(X,Y) = 1$ *if and only if there exist* $a > 0$ *and* $b \in \mathbb{R}$ *such that* $Y = aX+b$.

(iii) $\operatorname{Corr}(X,Y) = -1$ *if and only if there exist* $a < 0$ *and* $b \in \mathbb{R}$ *such that* $Y = aX + b$.

We postpone the proof of Theorem 4.37 and first look at an example.

Example 4.38. (Continuing Example 4.32 of multinomial variables.) In Example 4.32 we deduced

$$\operatorname{Cov}(X_i, X_j) = \begin{cases} np_i(1-p_i), & i = j \\ -np_ip_j, & i \ne j. \end{cases}$$

Thus, for $i \ne j$ we have

$$\begin{aligned}\operatorname{Corr}(X_i,X_j) &= \frac{\operatorname{Cov}(X_i,X_j)}{\sqrt{\operatorname{Var}(X_i)}\sqrt{\operatorname{Var}(X_j)}} = \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}} \\ &= -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.\end{aligned}$$

Notice that the correlation removes the factor n that measures the number of trials.

Consider the case $r = 2$. This is the binomial case, with $p_2 = 1 - p_1$, and we get

$$\operatorname{Corr}(X_1,X_2) = -\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}} = -1.$$

This reflects the linear relationship $X_2 = n - X_1$ of the binomial case. △

To prove Theorem 4.37, we introduce one of the most famous inequalities of mathematics.

Theorem 4.39 (Cauchy-Schwarz inequality)**.** *Let* X *and* Y *be two random variables. Assume that the expectations* $E(X^2)$*,* $E(Y^2)$ *and* $E(XY)$ *are finite. Then*

$$|\,E(XY)\,| \le \sqrt{E(X^2)}\sqrt{E(Y^2)}. \tag{4.30}$$

Equality holds iff there is a real number s *such that either* $P(X = sY) = 1$ *or* $P(Y = sX) = 1$.

With some additional work the assumption that the left-hand side of (4.30) is finite can be dropped (Corollary 4.40 below). This inequality appears in several parts of mathematics in various guises. For example, for any two real vectors $\mathbf{x} = (x_1, \dots, x_N)$ and $\mathbf{y} = (y_1, \dots, y_N)$,

$$\Big| \sum_{i=1}^{N} x_i y_i \Big| \le \Big(\sum_{i=1}^{N} x_i^2 \Big)^{1/2} \Big(\sum_{i=1}^{N} y_i^2 \Big)^{1/2}.$$

Equivalently, $\mathbf{x} \cdot \mathbf{y} \le |\mathbf{x}|\,|\mathbf{y}|$ where $|\mathbf{x}|$ denotes the Euclidean norm of $\mathbf{x}$. Also, for any two real functions that can be integrated,

$$\Big| \int_a^b f(x) g(x)\, dx \Big| \le \Big(\int_a^b f(x)^2\, dx \Big)^{1/2} \Big(\int_a^b g(x)^2\, dx \Big)^{1/2}.$$

The vectors and functions can also be complex valued. In the complex case the squares x_i^2 are replaced by $|x_i|^2$.

Proof of Theorem 4.39. First we take care of an uninteresting case.

Case 1. Suppose $E(Y^2) = 0$. Then Theorem 4.14(iv) implies $P(Y^2 = 0) = 1$. Since $Y^2 = 0$ implies $Y = 0$ implies $XY = 0$, we have $1 = P(Y^2 = 0) = P(Y = 0) = P(XY = 0)$. Consequently also $E(XY) = 0$ and so both sides of inequality (4.30) are zero. Thus the inequality is true in a trivial sense. Furthermore, the criterion for equality holds because now $Y = 0 \cdot X$. So the entire theorem is true.

Case 2. Suppose $E(Y^2) > 0$. For real t, let

$$f(t) = E[(X - tY)^2] = E(X^2) - 2tE(XY) + t^2 E(Y^2).$$

Since expectation of a square is nonnegative, $f(t) \ge 0$ for all t. Let us find the minimum of f.

$$f'(t) = 2tE(Y^2) - 2E(XY).$$
$$f''(t) = 2E(Y^2) > 0.$$

Hence f has a minimum at t_* defined by

$$f'(t_*) = 0 \iff t_* = \frac{E(XY)}{E(Y^2)}.$$

The minimum satisfies

$$\begin{aligned} 0 \le f(t_*) &= E(X^2) - 2t_* E(XY) + t_*^2 E(Y^2) \\ &= E(X^2) - 2\frac{E(XY)}{E(Y^2)} E(XY) + \Big(\frac{E(XY)}{E(Y^2)} \Big)^2 E(Y^2) \\ &= E(X^2) - \frac{\{E(XY)\}^2}{E(Y^2)}. \end{aligned}$$

Rearranging gives $\{E(XY)\}^2 \le E(X^2)E(Y^2)$, and taking square roots gives (4.30).

We investigate the case for equality in (4.30). Suppose first that $X = sY$ with probability 1. Then the left-hand side of (4.30) equals

$$|E(sY \cdot Y)| = |s| E(Y^2)$$

while the right-hand side equals

$$\sqrt{E(s^2Y^2)}\sqrt{E(Y^2)} = |s|E(Y^2).$$

Thus the inequality reduces to an equality. A similar argument works if $Y = sX$ with probability 1.

Next, suppose we have equality in (4.30). This gives $\{E(XY)\}^2 = E(X^2)E(Y^2)$ and hence $f(t_*) = 0$. From this

$$0 = f(t_*) = E[(X - t_*Y)^2].$$

By Theorem 4.14(iv), $P(X = t_*Y) = 1$. Proof of the theorem is complete. □

As an application, we prove the properties of the correlation coefficient.

Proof of Theorem 4.37. Part (i). Observe these equivalences:

$$\begin{aligned} &|\rho(X,Y)| \le 1 \\ \Longleftrightarrow\ &|\mathrm{Cov}(X,Y)| \le \sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)} \\ \Longleftrightarrow\ &|\,E[(X-\mu_X)(Y-\mu_Y)]\,| \le \sqrt{E[(X-\mu_X)^2]}\,\sqrt{E[(Y-\mu_Y)^2]}. \end{aligned} \tag{4.31}$$

The last statement is exactly Cauchy-Schwarz applied to the variables $X - \mu_X$ and $Y - \mu_Y$.

Parts (ii) and (iii). We investigate the case of equality. First calculate as follows. If $Y = aX + b$ with $a \neq 0$ then

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, aX+b) = a\,\mathrm{Cov}(X,X) = a\,\mathrm{Var}(X)$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X),$$

and so

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} = \frac{a\,\mathrm{Var}(X)}{|a|\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(X)}} = \frac{a}{|a|}. \tag{4.32}$$

We see that $\rho(X,Y) = \pm 1$ accordingly as $a/|a| = \pm 1$.

Now suppose $\rho(X,Y) = 1$ or -1. Then there is equality above in (4.31), in the Cauchy-Schwarz inequality. By Theorem 4.39 we have one of these two situations with probability 1, for some real s:

$$X - \mu_X = s(Y - \mu_Y) \quad \text{or} \quad Y - \mu_Y = s(X - \mu_X).$$

We cannot have $s = 0$ because then either X or Y would be constant with probability 1 and therefore have zero variance, which has been ruled out by assumption. Hence $s \neq 0$. By dividing by s if necessary, we get an identity of the form $Y = aX + b$ for $a = s$ or $1/s$ and for some real b.

It remains to check that a has the same sign as $\rho(X,Y)$. Calculation 4.32 above applies again and gives

$$\pm 1 = \rho(X,Y) = \frac{a}{|a|}$$

which forces a to have the same sign as $\rho(X,Y)$. The proof is complete. □

The assumption that $E(XY)$ is finite is actually superfluous in the Cauchy-Schwarz inequality of Theorem 4.39, as stated in the next corollary. We leave its proof as Exercise 4.27. The proof is an application of the monotone convergence theorem (Theorem 4.54) from Section 4.3.

Corollary 4.40. *Let X and Y be two random variables, and assume that both $E(X^2)$ and $E(Y^2)$ are finite. Then $E(XY)$ is also finite. In particular, Theorem 4.39 holds without assuming ahead of time that $E(XY)$ is finite.*

4.3. Construction of the expectation♣

This section develops the general definition and properties of the expectation to the point where we can prove the basic formulas for expectations of discrete and continuous random variables. We start from the beginning, so the discussion of Section 4.1 is not assumed.

A discrete random variable that takes only finitely many different real values is called a *simple random variable.* If X is simple, let $\alpha_1, \dots, \alpha_m$ be its distinct values. Then the events $A_i = \{X = \alpha_i\}$ form a partition of Ω. Using indicator variables

$$I_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}$$

we can express X as a linear combination of indicators:

$$X(\omega) = \sum_{i=1}^{m} \alpha_i \, I_{A_i}(\omega). \tag{4.33}$$

Assume that there a fixed probability space $(\Omega, \mathcal{F}, P)$ on which all the random variables discussed below are defined. The expectation EX (altenative notations: $E(X)$, $E[X]$, $E\{X\}$) of a random variable X is defined in three stages:

Step 1. Nonnegative simple X.

Step 2. $[0, \infty]$-valued X.

Step 3. $[-\infty, \infty]$-valued X.

Next we take each step slowly in turn.

Step 1. Let X be a nonnegative simple random variable with distinct values $\alpha_1, \dots, \alpha_m$. Its expectation EX is defined by

$$EX = \sum_{i=1}^{m} \alpha_i P(A_i) = \sum_{i=1}^{m} \alpha_i P(X = \alpha_i). \tag{4.34}$$

If $\alpha_i = 0$ for some i, including it or leaving it out from the sum above makes no difference to EX. △

Example 4.41. Let $\Omega = [0, 1]$ with probability given by length: $P((a, b)) = P([a, b]) = b - a$ for $[a, b] \subset [0, 1]$. Define X by

$$X(\omega) = \begin{cases} 1, & 0 \le \omega < \frac{1}{\pi} \\ e, & \frac{1}{\pi} \le \omega \le \frac{1}{2} \\ 10, & \frac{1}{2} < \omega \le 1. \end{cases}$$

Then

$$EX = 1 \cdot P([0, \tfrac{1}{\pi})) + e \cdot P([\tfrac{1}{\pi}, \tfrac{1}{2}]) + 10 \cdot P((\tfrac{1}{2}, 1]) = \tfrac{1}{\pi} + e \cdot (\tfrac{1}{2} - \tfrac{1}{\pi}) + 5.$$

△

Before taking the next step, we check that this definition has certain desirable properties. First a small extension of the definition.

Lemma 4.42. *Let $\{B_j\}_{j=1}^n$ be a partition of Ω, $\{\beta_j\}_{j=1}^n$ nonnegative reals, and $X(\omega) = \sum_{j=1}^n \beta_j I_{B_j}(\omega)$. Then $EX = \sum_{j=1}^n \beta_j P(B_j)$.*

Proof. If the numbers β_j are distinct, then this is a restatement of the definition. However, now some β_js may coincide. Let $\alpha_1, \dots, \alpha_m$ be the distinct numbers among the $\beta_1, \dots, \beta_n$ (and then $m \le n$). Set $J(i) = \{j : \beta_j = \alpha_i\}$ and define events $A_i = \bigcup_{j \in J(i)} B_j$. Now $\alpha_1, \dots, \alpha_m$ cover at least all the distinct nonzero values of X, and $A_i = \{X = \alpha_i\}$. By definition (4.34) and some rearrangement,

$$\begin{aligned} EX &= \sum_{i=1}^m \alpha_i P(A_i) = \sum_{i=1}^m \alpha_i \sum_{j=1}^n I_{j \in J(i)}\, P(B_j) = \sum_{j=1}^n P(B_j) \sum_{i=1}^m \alpha_i\, I_{j \in J(i)} \\ &= \sum_{j=1}^n \beta_j P(B_j). \end{aligned}$$

The last step follows because for a given j, the condition $j \in J(i)$ picks the unique i that satisfies $\alpha_i = \beta_j$. □

Lemma 4.43. *Let X and Y be nonnegative simple random variables. Then their expectations have these properties.*

Linearity: for real $a, b \ge 0$, $E[aX + bY] = aEX + bEY$.

Monotonicity: if $X \ge Y$, then $EX \ge EY$.

Proof. Let X and Y have representations

$$X = \sum_{i=1}^m \alpha_i\, I_{A_i} \quad \text{and} \quad Y = \sum_{j=1}^n \beta_j\, I_{B_j}$$

with partitions $\{A_i\}$ and $\{B_j\}$. Since each A_i partitions as $A_i = \bigcup_j (A_i \cap B_j)$, and similarly $B_j = \bigcup_i (A_i \cap B_j)$, we can rewrite X and Y as

$$X = \sum_{i,j} \alpha_i\, I_{A_i \cap B_j} \quad \text{and} \quad Y = \sum_{i,j} \beta_j\, I_{A_i \cap B_j} \tag{4.35}$$

so that they use the same partition $\{A_i \cap B_j\}_{1 \le i \le m,\, 1 \le j \le n}$. Then the linear combination can be written as

$$aX + bY = \sum_{i,j} a\alpha_i\, I_{A_i \cap B_j} + \sum_{i,j} b\beta_j\, I_{A_i \cap B_j} = \sum_{i,j} (a\alpha_i + b\beta_j)\, I_{A_i \cap B_j}.$$

Apply Lemma 4.42 in the second equality below:

$$\begin{aligned} E[aX+bY] &= E\Bigl[\sum_{i,j}(a\alpha_i+b\beta_j)\,I_{A_i\cap B_j}\Bigr] = \sum_{i,j}(a\alpha_i+b\beta_j)\,P(A_i\cap B_j) \\ &= a\sum_{i,j}\alpha_i\,P(A_i\cap B_j)+b\sum_{i,j}\beta_j\,P(A_i\cap B_j) = a\sum_i\alpha_i\,P(A_i)+b\sum_j\beta_j\,P(B_j) \\ &= aEX+bEY. \end{aligned}$$

Some of the sets $A_i\cap B_j$ may be empty. These make no difference to the calculation because then $I_{A_i\cap B_j}=0$ and $P(A_i\cap B_j)=0$.

Under the monotonicity assumption $X\ge Y$, representation (4.35) implies that $\alpha_i\ge\beta_j$ whenever $A_i\cap B_j\ne\varnothing$ (because $X(\omega)\ge Y(\omega)$ for each $\omega\in A_i\cap B_j$). Then it follows that each term of the sum

$$EX=\sum_{i,j}\alpha_i\,P(A_i\cap B_j)$$

is at least as large as the corresponding term of the sum

$$EY=\sum_{i,j}\beta_j\,P(A_i\cap B_j).$$

This gives $EX\ge EY$. □

Linearity extends inductively to any finite number of terms: Suppose $X_1,\dots,X_n$ are nonnegative simple random variables and $a_1,\dots,a_n$ are nonnegative reals. Then any linear combination $\sum_{i=1}^k a_iX_i$ is also a nonnegative simple random variable, and we can use Lemma 4.43 repeatedly:

$$\begin{aligned} E\Bigl[\sum_{i=1}^n a_iX_i\Bigr] &= E\Bigl[\sum_{i=1}^{n-1} a_iX_i+a_nX_n\Bigr] = E\Bigl[\sum_{i=1}^{n-1} a_iX_i\Bigr]+a_nEX_n \\ &= \cdots = \sum_{i=1}^n a_i\,EX_i. \end{aligned}$$

We take the next step in the definition of the expectation.

Step 2. Let X be a $[0,\infty]$-valued random variable. Then its expectation is defined by

$$\begin{aligned} EX = \sup\{EW : \ &W \text{ is a simple random variable on } \Omega \text{ such that} \\ &0\le W(\omega)\le X(\omega) \text{ for all } \omega\in\Omega\}. \end{aligned} \tag{4.36}$$

This gives a value $EX\in[0,\infty]$. △

In case X is also simple, we need to check that (4.36) gives the same value to EX as Step 1. This is true because $W\le X$ implies $EW\le EX$ by Lemma 4.43, so the supremum is attained at $W=X$.

The next example shows that the expectation can be infinite even though the random variable takes only finite nonnegative values.

Example 4.44. Let $\Omega = \mathbb{Z}_{>0} = \{1, 2, 3, \dots\}$. Define P by $P\{\omega\} = \frac{c}{\omega^2}$ for $\omega \in \Omega$ where $c = (\sum_{k\ge1} \frac{1}{k^2})^{-1} = (\pi^2/6)^{-1}$. Let $X(\omega) = \omega$ on Ω. Define simple random variables

$$Y_n(\omega) = \sum_{k=1}^{n} k\, I_k(\omega) = \begin{cases} \omega, & 1 \le \omega \le n \\ 0, & \omega > n. \end{cases}$$

$Y_n \le X$ and $EY_n = c\sum_{k=1}^n k \cdot k^{-2} = c\sum_{k=1}^n k^{-1} \to \infty$ as $n \to \infty$. Hence $EX = \infty$. $\triangle$

Remark 4.45. If $P(X = \infty) > 0$ then $EX = \infty$ because for each $M > 0$, the simple random variable $X_M = M \cdot I_{\{X=\infty\}}$ satisfies both $0 \le X_M \le X$ and

$$EX \ge E[X_M] = M \cdot P(X = \infty).$$

We can let $M \nearrow \infty$ on the right to conclude that $EX = \infty$.

The converse statement is useful enough to warrant stating as the next theorem. $\triangle$

Theorem 4.46. *Let X be a $[0, \infty]$-valued random variable such that $EX < \infty$. Then $P(X < \infty) = 1$.*

Example 4.47. Here is an example of a naturally arising random variable with $P(X = \infty) > 0$. Imagine a robot that shoots basketball hoops. After every shot, the accuracy of the robot improves. Precisely speaking, assume that for $k = 1, 2, 3, \dots$, the kth shot succeeds with probability $\exp(-\frac{1}{k^2})$, independently of the other shots.

Let X be the number of shots taken when the first miss happens. Set $X = \infty$ if the robot never misses. Let us calculate the probability that $X = \infty$. Let A_k be the event that the kth shot succeeds. Use below the independence of the events A_k.

$$\begin{aligned} P(X = \infty) &= P\Big(\bigcap_{k=1}^{\infty} A_k\Big) = \lim_{n\to\infty} P\Big(\bigcap_{k=1}^{n} A_k\Big) = \lim_{n\to\infty} \prod_{k=1}^{n} P(A_k) \\ &= \lim_{n\to\infty} \prod_{k=1}^{n} \exp(-\tfrac{1}{k^2}) = \lim_{n\to\infty} \exp\Big(-\sum_{k=1}^{n} \tfrac{1}{k^2}\Big) = \exp\Big(-\sum_{k=1}^{\infty} \tfrac{1}{k^2}\Big) \\ &= \exp(-\pi^2/6). \end{aligned}$$

$\triangle$

We also check that altering a random variable on an event of probability zero cannot change the expectation.

Lemma 4.48. *Suppose X and $\widetilde{X}$ are two $[0, \infty]$-valued random variables that agree with probability one, that is, $P(X = \widetilde{X}) = 1$. Then $EX = E\widetilde{X}$.*

Proof. The numbers EX and $E\widetilde{X}$ are by definition the suprema $EX = \sup\mathcal{U}$ and $E\widetilde{X} = \sup\widetilde{\mathcal{U}}$ for the sets of real numbers

$$\mathcal{U} = \{EW : 0 \le W \le X,\ W \text{ is a simple r.v.}\}$$
$$\text{and}\quad \widetilde{\mathcal{U}} = \{E\widetilde{W} : 0 \le \widetilde{W} \le \widetilde{X},\ \widetilde{W} \text{ is a simple r.v.}\}.$$

We show that $E\widetilde{X}$ is an upper bound of $\mathcal{U}$. By the definition of the supremum, this implies that $EX \le E\widetilde{X}$. Since the roles of X and $\widetilde{X}$ can be switched around, we also have $E\widetilde{X} \le EX$ and therefore equality $EX = E\widetilde{X}$.

So suppose $W = \sum_{j=1}^{m} \beta_j \, I_{B_j}$ is a simple random variable such that $0 \le W \le X$. Let $A = \{X = \widetilde{X}\}$. Then $Z = W I_A = \sum_{j=1}^{m} \beta_j \, I_{B_j \cap A}$ is also a nonnegative simple random variable. $P(A) = 1$ implies $P(B_j \cap A) = P(B_j)$ and hence the expectations of Z and W agree: $EZ = EW$. But now

$$\begin{aligned} \text{for } \omega \in A, \quad & Z(\omega) = W(\omega) \le X(\omega) = \widetilde{X}(\omega) \\ \text{while for } \omega \in A^c, \quad & Z(\omega) = 0 \le \widetilde{X}(\omega). \end{aligned}$$

Thus $Z \le \widetilde{X}$, and we conclude that $E\widetilde{X} \ge EZ = EW$. This works for any number $EW \in \mathcal{U}$, and therefore $E\widetilde{X}$ is an upper bound for $\mathcal{U}$. □

The lemma above takes care of examples such as the following, which is actually a very standard way we might construct a random variable.

Example 4.49. Let $\Omega = \mathbb{R}_+ = [0, \infty)$, the nonnegative real line. Let $0 < p < 1$. Define the probability measure P on Ω by setting $P\{0\} = 1 - p$, $P\{1\} = p$, and $P(B) = 0$ for any subset $B \subset \Omega$ that does not intersect $\{0, 1\}$. Let $X : \Omega \to \mathbb{R}$ be the identity function: $X(\omega) = \omega$ for $0 \le \omega < \infty$. Then note that X is a Ber(p) random variable because $P(X = 0) = P\{0\} = 1 - p$ and $P(X = 1) = P\{1\} = p$, but X is not simple because it takes uncountably many values!

However, if we define $Y : \Omega \to \mathbb{R}$ by

$$Y(\omega) = \begin{cases} 0, & 0 \le \omega < 1 \\ 1, & \omega \ge 1 \end{cases}$$

then Y is simple, $EY = 0 \cdot P([0, 1)) + 1 \cdot P([1, \infty)) = p$, and $P(X = Y) = P\{0, 1\} = 1$. Consequently we can conclude that $EX = EY = p$. △

It is very convenient for examples to know that instead of the forbidding supremum in definition (4.36), we can take any sequence of simple functions that converge monotonically to X. This is the content of the next lemma whose proof is unfortunately rather technical.

Lemma 4.50. *Let X be a $[0, \infty]$-valued random variable. Let Y_n be a sequence of simple random variables such that $0 \le Y_n \nearrow X$. By this we mean that $0 \le Y_1(\omega) \le Y_2(\omega) \le Y_3(\omega) \le \cdots \le X(\omega)$ and $\lim_{n\to\infty} Y_n(\omega) = X(\omega)$, for all $\omega \in \Omega$. Then $EY_n \to EX$.*

Proof. The sequence of numbers EY_n is monotone, so the limit $c = \lim_{n\to\infty} EY_n$ exists in $[0, \infty]$. Our goal is to show that $c = EX$. This will come in two steps: first $EX \ge c$ and then $c \ge EX$.

The definition (4.36) of EX implies that $EX \ge EY_n$ for all n, and so $EX \ge c$.

To get the opposite inequality $c \ge EX$, we have to rely on the definition (4.36) of EX and consider an arbitrary simple random variable $W = \sum_{j=1}^{m} \beta_j \, I_{B_j}$ such that $\{B_j\}$ is a partition of Ω and $0 \le W \le X$. Let also $\varepsilon > 0$ be an arbitrary small positive real. Our first intermediate goal is to show that

$$c \ge EW - \varepsilon. \tag{4.37}$$

After that a small analysis argument takes us to the finish line.

Define events $A_n = \{Y_n \ge W - \varepsilon\}$. These events are increasing by the monotonicity of the sequence Y_n, that is, $A_1 \subset A_2 \subset A_3 \subset \cdots$. (Justification: $\omega \in A_n$ is the same as $Y_n(\omega) \ge W(\omega) - \varepsilon$, which implies $Y_{n+1}(\omega) \ge Y_n(\omega) \ge W(\omega) - \varepsilon$, which says $\omega \in A_{n+1}$.)

We claim that $\bigcup_n A_n = \Omega$. Given any ω, the limit assumption $Y_n(\omega) \to X(\omega)$ and $X(\omega) \ge W(\omega) > W(\omega) - \varepsilon$ imply that, for some n, $Y_n(\omega) \ge W(\omega) - \varepsilon$, so that $\omega \in A_n$. So each ω lies in some A_n. This proves the claim.

On the event A_n, $Y_n \ge W - \varepsilon$. $W - \varepsilon$ is a simple random variable, but it might fail to be nonnegative, so its integral is not yet defined. We define another simple random variable $\widetilde{W}$ that equals $W - \varepsilon$ whenever this latter is nonnegative, and equals zero if $W - \varepsilon < 0$. Precisely speaking, first put

$$\widetilde{\beta}_j = (\beta_j - \varepsilon)^+ = \begin{cases} \beta_j - \varepsilon, & \text{if } \beta_j - \varepsilon \ge 0 \\ 0, & \text{if } \beta_j - \varepsilon < 0 \end{cases}$$

and then

$$\widetilde{W} = \sum_{j=1}^{m} \widetilde{\beta}_j \, I_{B_j}.$$

We claim that on the event A_n, we have $Y_n \ge \widetilde{W}$. Justification: Let $\omega \in A_n$. Pick j such that $\omega \in B_j$. Then $W(\omega) = \beta_j$. There are two cases for $\widetilde{\beta}_j$:

(i) If $\widetilde{\beta}_j = \beta_j - \varepsilon$, then from $\omega \in A_n$ we reason that $Y_n(\omega) \ge W(\omega) - \varepsilon = \beta_j - \varepsilon = \widetilde{\beta}_j = \widetilde{W}(\omega)$.

(ii) If $\widetilde{\beta}_j = 0$, then from the nonnegativity of Y_n we reason that $Y_n(\omega) \ge 0 = \widetilde{\beta}_j = \widetilde{W}(\omega)$.

From these two cases we conclude that $Y_n \ge \widetilde{W}$ on the event A_n. This has the consequence that

$$Y_n(\omega) I_{A_n}(\omega) \ge \widetilde{W}(\omega) I_{A_n}(\omega) \qquad \text{for all } \omega \in \Omega$$

because for $\omega \notin A_n$ this inequality reduces to $0 \ge 0$.

Note that multiplying a nonnegative simple random variable with an indicator random variable results in another nonnegative simple random variable. Note also that a product of indicators is the indicator of the intersection: $I_A I_B = I_{A \cap B}$.

Now using Lemma 4.43 and what was proved above,

$$EY_n \ge E[Y_n I_{A_n}] \ge E\big[\widetilde{W} I_{A_n}\big] = E\biggl[\sum_{j=1}^{m} \widetilde{\beta}_j \, I_{B_j \cap A_n}\biggr] = \sum_{j=1}^{m} \widetilde{\beta}_j \, P(B_j \cap A_n).$$

Next we let $n \to \infty$. We argued above that $A_n \nearrow \Omega$. Hence also $B_j \cap A_n \nearrow B_j$ as $n \to \infty$. So by the continuity of probability, $P(B_j \cap A_n) \to P(B_j)$ as $n \to \infty$.

From the definition of $\widetilde{\beta}_j$ we have the inequality $\widetilde{\beta}_j \ge \beta_j - \varepsilon$. Using all this we arrive at the following:

$$c = \lim_{n\to\infty} EY_n \ge \lim_{n\to\infty} \sum_{j=1}^m \widetilde{\beta}_j \, P(B_j \cap A_n) = \sum_{j=1}^m \widetilde{\beta}_j \, P(B_j) \ge \sum_{j=1}^m (\beta_j - \varepsilon) \, P(B_j)$$

$$= \sum_{j=1}^m \beta_j \, P(B_j) - \varepsilon \sum_{j=1}^m P(B_j) = EW - \varepsilon.$$

We have reached (4.37), namely that $c \ge EW - \varepsilon$. The key point is that this inequality is valid for all $\varepsilon > 0$ and for all nonnegative simple random variables W that satisfy $0 \le W \le X$, in other words, all those that appear on the right-hand side of definition (4.36).

Next we use this point from analysis: if $a \ge b - \varepsilon$ for all $\varepsilon > 0$, then $a \ge b$. Applied to $c \ge EW - \varepsilon$ for all $\varepsilon > 0$, we conclude that

$$c \ge EW \qquad \text{for all } W \text{ in definition (4.36).}$$

This tells us that c is an upper bound of the set

$$\{EW : 0 \le W \le X,\ W \text{ is a simple random variable on } \Omega\}.$$

Hence the supremum of this set, namely EX, satisfies $EX \le c$. The proof is complete. □

The expectation of a nonnegative discrete random variable with finitely many values is covered by the definition in Step 1. With the lemma above we can derive the expectation of a nonnegative discrete random variable with countably infinitely many values.

Theorem 4.51. *Let $0 \le x_1 < x_2 < x_3 < \cdots$ be an increasing sequence of nonnegative reals. Let X be a nonnegative random variable such that $\sum_{k=1}^\infty P(X = x_k) = 1$. Then EX is a well-defined value in $[0,\infty]$ given by $EX = \sum_{k=1}^\infty x_k \, P(X = x_k)$.*

Proof. First we use Lemma 4.48 to replace X with a random variable whose range is countable. Let

$$A = \{\omega : X(\omega) \in \{x_1, x_2, x_3, \dots\}\} = \bigcup_{k=1}^\infty \{X = x_k\}$$

be the event on which X takes one of the values x_k. We do not know what X does on A^c, but the assumption is that $P(A^c) = 0$. Hence if we define

$$Y(\omega) = X(\omega) I_A(\omega) = \begin{cases} X(\omega) & \text{if } \omega \in A, \\ 0 & \text{if } \omega \in A^c, \end{cases}$$

we have a random variable Y with countable range $\{0, x_1, x_2, x_3, \dots\}$ such that $P(X = Y) \ge P(A) = 1$. (This is true regardless of whether $x_1 = 0$ which is possible.) By Lemma 4.48 $EX = EY$, so now it remains to calculate EY.

Define the sequence of nonnegative simple functions

$$Y_n(\omega) = \sum_{k=1}^n x_k \, I_{\{X = x_k\}}(\omega)$$

This is a monotone nondecreasing sequence:

$$Y_{n+1} - Y_n = x_{n+1}\, I_{\{X=x_{n+1}\}} \ge 0.$$

To argue that $Y_n(\omega) \nearrow Y(\omega)$ for each ω, we have cases.

(i) If $\omega \in A^c$ then $Y_n(\omega) = 0 = Y(\omega)$ for all n.

(ii) If $\omega \in A$ then for some m, $\omega \in \{X = x_m\}$, and consequently for $n \ge m$ we have $Y_n(\omega) = x_m = Y(\omega)$.

Lemma 4.50 now tells us that

$$EY = \lim_{n\to\infty} EY_n = \lim_{n\to\infty} \sum_{k=1}^{n} x_k\, P(X = x_k) = \sum_{k=1}^{\infty} x_k\, P(X = x_k).$$

The last equality is simply the *definition* of the series $\sum_{k=1}^{\infty} x_k\, P(X = x_k)$ as the limit of its partial sums. Since the terms $x_k\, P(X = x_k)$ are nonnegative, the partial sums form a monotone nondecreasing sequence in $[0, \infty)$. Such as a sequence is either bounded in which case it converges to a finite limit, or it is unbounded, in which case the limit is taken to be ∞. □

It is also useful to have a general recipe that gives a concrete approximating sequence of simple random variables for any nonnegative random variable. For $n \in \mathbb{Z}_{>0}$ define the function $g_n : [0, \infty] \to [0, \infty)$ by

$$g_n(x) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\, I_{\{\frac{k}{2^n} < x \le \frac{k+1}{2^n}\}} + n\, I_{\{x>n\}} \tag{4.38}$$

Sketch the graph of g_n. The range of g_n is the finite set $\{0, \frac{1}{2^n}, \frac{2}{2^n}, \frac{3}{2^n}, \dots, n-\frac{1}{2^n}, n\}$, $g_n(0) = 0$, and for each $x \in (0, \infty]$, $g_n(x) < x$ and $g_n(x)$ is a nondecreasing sequence that converges to x. (Exercise 4.21 asks for the details.) This monotonicity is achieved by use of the dyadic partitions (partitions into intervals with endpoints $k/2^n$). Using these functions g_n gives us a particular approximating sequence of simple random variables.

Theorem 4.52. *Let X be a $[0, \infty]$-valued random variable. Then $g_n(X)$ is a nonnegative simple random variable, and*

$$EX = \lim_{n\to\infty} E[g_n(X)] = \lim_{n\to\infty} \left\{ \sum_{k=0}^{n2^n-1} \frac{k}{2^n} P(\tfrac{k}{2^n} < X \le \tfrac{k+1}{2^n}) + nP(X > n) \right\}. \tag{4.39}$$

Proof. The pointwise limit $0 \le g_n(x) \nearrow x$ for $x \in [0, \infty]$ gives $0 \le g_n(X(\omega)) \nearrow X(\omega)$. The limit $EX = \lim_{n\to\infty} E[g_n(X)]$ then comes from Lemma 4.50. The formula for $E[g_n(X)]$ comes from taking expectations term by term of the expression

$$g_n(X(\omega)) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} I_{\{\frac{k}{2^n} < X \le \frac{k+1}{2^n}\}}(\omega) + nI_{\{X>n\}}(\omega). \tag{4.40}$$

□

Here is the third and final step in the definition of the expectation.

Step 3. Let X be a $[-\infty, \infty]$-valued random variable on Ω. Its positive and negative parts $X^+(\omega) = X(\omega) \vee 0$ and $X^-(\omega) = (-X(\omega)) \vee 0$ are $[0, \infty]$-valued random variables. The expectations $E(X^+)$ and $E(X^-)$ were defined in Step 2. If at least one of them is finite, the expectation of X is defined by

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

If $E(X^+) = E(X^-) = \infty$, then EX is not defined. $\triangle$

A necessary and sufficient condition for having a finite mean EX is that $E|X| < \infty$. This is because $X^\pm \le |X| = X^+ + X^-$ and consequently $E[X^\pm]$ are both finite if and only if $E|X|$ is finite.

Theorem 4.53. *Let X and Y be either $[0, \infty]$-valued or such that their expectations are finite. Then their expectations have these properties.*

$$E[X+Y] = EX + EY. \tag{4.41}$$

$$E[aX] = a\,EX \quad \textit{for real } a. \tag{4.42}$$

$$\textit{If} \quad X \ge Y \quad \textit{then} \quad EX \ge EY. \tag{4.43}$$

$$\textit{If} \quad P(X = Y) = 1 \quad \textit{then} \quad EX = EY. \tag{4.44}$$

$$\textit{If} \quad X \overset{d}{=} Y \quad \textit{then} \quad EX = EY. \tag{4.45}$$

Proof. Suppose first $X, Y \ge 0$ and $a, b \ge 0$. Pick any simple random variables $0 \le X_n \nearrow X$ and $0 \le Y_n \nearrow Y$. Then we also have $0 \le aX_n + bY_n \nearrow aX + bY$. Consequently

$$\begin{aligned} E[aX + bY] &= \lim_{n\to\infty} E[aX_n + bY_n] = \lim_{n\to\infty} \bigl(aE[X_n] + bE[Y_n]\bigr) \\ &= a \lim_{n\to\infty} E[X_n] + b \lim_{n\to\infty} E[Y_n] = aEX + bEY. \end{aligned}$$

The second equality is the linearity of the expectation for simple random variables from Lemma 4.43. The third equality is the linearity of limits, which is true even if one or both limits are infinite, since everything is nonnegative now.

If $0 \le Y \le X$, then definition (4.36) shows $EY \le EX$ because each W on the right-hand side for Y is there also for X. Hence EX is the supremum of a larger set than EY.

Now assume X, Y real with finite expectations EX, EY. Let $W = X + Y$. From

$$W^+ - W^- = X^+ - X^- + Y^+ - Y^-$$

we get

$$W^+ + X^- + Y^- = W^- + X^+ + Y^+.$$

By the additivity of the expectation of nonnegative random variables,

$$EW^+ + EX^- + EY^- = EW^- + EX^+ + EY^+$$

which rearranges to

$$EW^+ - EW^- = EX^+ - EX^- + EY^+ - EY^-$$

which says exactly that $EW = EX + EY$, as desired. (4.41) is proved.

Next assume $a \geq 0$. Then $(aX)^{\pm} = aX^{\pm}$, and we reason as follows:

$$\begin{aligned}E[aX] &= E[(aX)^+] - E[(aX)^-] = E[aX^+] - E[aX^-] = aE[X^+] - aE[X^-] \\ &= a\big(E[X^+] - E[X^-]\big) = aEX.\end{aligned}$$

The third equality used the case of $a, X \geq 0$ proved above. Next, suppose $a < 0$. Then, using the step above for $-a > 0$ and $(-X)^{\pm} = X^{\mp}$,

$$\begin{aligned}E[aX] &= E[-a(-X)] = -aE[-X] = -a(E[X^-] - E[X^+]) \\ &= a(E[X^+] - E[X^-]) = aEX.\end{aligned}$$

(4.42) is proved.

Suppose $Y(\omega) \leq X(\omega)$ for all ω. It follows that $Y^+(\omega) \leq X^+(\omega)$ and $Y^-(\omega) \geq X^-(\omega)$ for all ω. Hence

$$EY = E(Y^+) - E(Y^-) \leq E(X^+) - E(X^-) = EX.$$

(4.43) is proved.

Let $A = \{X = Y\}$. The for any $\omega \in A$, we have $X^+(\omega) = X(\omega) \vee 0 = Y(\omega) \vee 0 = Y^+(\omega)$, and also $X^-(\omega) = (-X(\omega)) \vee 0 = (-Y(\omega)) \vee 0 = Y^-(\omega)$. Since $P(A) = 1$ by assumption, we have that $P(X^+ = Y^+) = P(X^- = Y^-) = 1$. Then by Lemma 4.48 applied to the pairs X^+, Y^+ and X^-, Y^-, we get

$$EX = E[X^+] - E[X^-] = E[Y^+] - E[Y^-] = EY.$$

(4.44) is proved.

Assume $X \overset{d}{=} Y$. We need to show that $E(X^+) = E(Y^+)$ and $E(X^-) = E(Y^-)$. Use Theorem 4.52 to express the expectations of X^+ and Y^+ as $E(X^+) = \lim_{n\to\infty} E[g_n(X^+)]$ and $E(Y^+) = \lim_{n\to\infty} E[g_n(Y^+)]$. The right-hand side of (4.39) shows that $E[g_n(X^+)] = E[g_n(Y^+)]$ because, by the assumption of equality in distribution, for all $k \geq 0$ and $n \geq 1$,

$$\begin{aligned}P(\tfrac{k}{2^n} < X^+ \leq \tfrac{k+1}{2^n}) &= P(\tfrac{k}{2^n} < X \leq \tfrac{k+1}{2^n}) = P(\tfrac{k}{2^n} < Y \leq \tfrac{k+1}{2^n}) \\ &= P(\tfrac{k}{2^n} < Y^+ \leq \tfrac{k+1}{2^n})\end{aligned}$$

and similarly $P(X^+ > n) = P(Y^+ > n)$. In the $n \to \infty$ limit $E[g_n(X^+)] = E[g_n(Y^+)]$ turns into $E(X^+) = E(Y^+)$. An analogous argument gives $E(X^-) = E(Y^-)$. (4.45) is proved. □

The point of the next theorem is that the converging sequence no longer necessarily consists of simple random variables. It is a very useful result for the theory.

Theorem 4.54 (Monotone convergence theorem). *Let X and a sequence X_n be $[0,\infty]$-valued random variables such that $0 \leq X_n \nearrow X$. By this we mean again that, for each $\omega \in \Omega$, $0 \leq X_1(\omega) \leq X_2(\omega) \leq X_3(\omega) \leq \cdots \leq X(\omega)$ and $\lim\limits_{n\to\infty} X_n(\omega) = X(\omega)$. Then $EX_n \to EX$.*

Proof. Let $Y_{n,k}$ be nonnegative simple random variables such that, for each n, $0 \leq Y_{n,k} \nearrow X_n$ as $k \to \infty$. Let $W_n = Y_{1,n} \vee Y_{2,n} \vee \cdots \vee Y_{n,n}$. Then W_n is also a nonnegative simple random variable, and the sequence is monotone: $0 \leq W_1 \leq W_2 \leq \cdots$.

We claim that $\lim_{n\to\infty} W_n = X$. First from the definition of W_n,

$$W_n = Y_{1,n} \vee \cdots \vee Y_{n,n} \;\le\; X_1 \vee \cdots \vee X_n = X_n \le X, \tag{4.46}$$

from which $\lim_{n\to\infty} W_n \le X$.

For any $n \ge m$ we have $W_n \ge Y_{m,n}$. Hence

$$\lim_{n\to\infty} W_n \ge \lim_{n\to\infty} Y_{m,n} = X_m.$$

Since this is true for all m, we can let $m \to \infty$ on the right to get $\lim_{n\to\infty} W_n \ge X$.

We have now checked that $0 \le W_n \nearrow X$ for simple random variables W_n. This implies $EW_n \to EX$. Equation (4.46) gives $EW_n \le EX_n \le EX$. Since EX_n is sandwiched between EW_n and EX, we conclude that $EX_n \to EX$. □

Theorem 4.55. *Suppose random variable X has probability density function f. Then*

$$E[X^+] = \int_0^\infty x\,f(x)\,dx \quad \textit{and} \quad E[X^-] = -\int_{-\infty}^0 x\,f(x)\,dx. \tag{4.47}$$

with values in $[0,\infty]$. Furthermore, if EX is well-defined, then

$$EX = \int_{-\infty}^\infty x\,f(x)\,dx. \tag{4.48}$$

Proof. We derive the expectation first for a truncated random variable. For each positive integer M, define

$$X_M(\omega) = X(\omega) I_{\{|X|\le M\}}(\omega) = \begin{cases} X(\omega) & \text{if } |X(\omega)| \le M \\ 0 & \text{if } |X(\omega)| > M. \end{cases}$$

Define nonnegative simple random variables

$$Y_{M,n}(\omega) = \sum_{k=0}^{M2^n-1} \frac{k}{2^n} I_{\{\frac{k}{2^n} < X \le \frac{k+1}{2^n}\}}(\omega).$$

Then $0 \le Y_{M,n} \nearrow X_M^+$ as $n \to \infty$, and hence

$$\begin{aligned} E[X_M^+] &= \lim_{n\to\infty} E[Y_{M,n}] = \lim_{n\to\infty} \sum_{k=0}^{M2^n-1} \frac{k}{2^n} P\big(\tfrac{k}{2^n} < X \le \tfrac{k+1}{2^n}\big) \\ &= \lim_{n\to\infty} \sum_{k=0}^{M2^n-1} \frac{k}{2^n} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} f(x)\,dx = \lim_{n\to\infty} \sum_{k=0}^{M2^n-1} \frac{k}{2^n} \int_0^M I_{\{\frac{k}{2^n} < x \le \frac{k+1}{2^n}\}}\, f(x)\,dx \\ &= \lim_{n\to\infty} \int_0^M f(x) \sum_{k=0}^{M2^n-1} \frac{k}{2^n} I_{\{\frac{k}{2^n} < x \le \frac{k+1}{2^n}\}}\,dx \\ &= \int_0^M x\,f(x)\,dx \;-\; \lim_{n\to\infty} \int_0^M f(x) \sum_{k=0}^{M2^n-1} \big(x - \tfrac{k}{2^n}\big) I_{\{\frac{k}{2^n} < x \le \frac{k+1}{2^n}\}}\,dx. \end{aligned}$$

The error term at the end satisfies these bounds:

$$0 \le \int_0^M f(x) \sum_{k=0}^{M2^n-1} \bigl(x - \tfrac{k}{2^n}\bigr) I_{\{\frac{k}{2^n} < x \le \frac{k+1}{2^n}\}}\, dx \;\le\; \frac{1}{2^n} \int_0^M f(x)\, dx \;\le\; \frac{1}{2^n}$$

and hence vanishes as $n \to \infty$.

We have verified

$$E[X_M^+] = \int_0^M x\, f(x)\, dx. \tag{4.49}$$

By the monotone limit $0 \le X_M^+ \nearrow X^+$ and the monotone convergence theorem,

$$E[X^+] = \lim_{M\to\infty} E[X_M^+] = \lim_{M\to\infty} \int_0^M x\, f(x)\, dx = \int_0^\infty x\, f(x)\, dx. \tag{4.50}$$

The last equality is the definition of the improper (Riemann) integral on the right, with a value in $[0,\infty]$. This proves the first part of (4.47).

Next let $Y = -X$. Then Y has density function $g(x) = f(-x)$. Here is the check. Change variables below with $y = -x$.

$$\begin{aligned} P(Y \le b) = P(X \ge -b) &= \int_{-b}^\infty f(x)\, dx = -\int_b^{-\infty} f(-y)\, dy = \int_{-\infty}^b f(-y)\, dy \\ &= \int_{-\infty}^b g(y)\, dy. \end{aligned}$$

Then line above shows that Y has density function $g(x) = f(-x)$.

By an application of (4.50) to Y, we have

$$E[Y^+] = \int_0^\infty y\, g(y)\, dy = \int_0^\infty y\, f(-y)\, dy.$$

Since $X^- = Y^+$, we have, with another change of variable $x = -y$:

$$E[X^-] = E[Y^+] = \int_0^\infty y\, f(-y)\, dy = \int_0^{-\infty} x\, f(x)\, dx = -\int_{-\infty}^0 x\, f(x)\, dx.$$

This proves also the second part of (4.47). (4.48) follows from (4.47) and $EX = E[X^+] - E[X^-]$. □

Justification of the results of Section 4.1.

We now explain how the results presented in Section 4.1 are justified from the theory developed in the present section. Theorem 4.5 comes from Theorem 4.51. Theorem 4.6 comes from Theorem 4.55. Theorem 4.8 is statement (4.45). Theorem 4.11 is Exercise 4.22. Theorem 4.25 on the product of expectations is Exercise 4.23.

Integration.

The development presented in this section is the development of the Lebesgue integral of measure theory, specialized to the case of a probability measure. The only point left unmentioned is that the random variables have to be assumed *measurable functions*. This means that all the events in this section, beginning with those of type $\{X = \alpha\}$ in (4.33), are members of $\mathcal{F}$.

We close this section by observing that the Riemann integral of calculus is a special case of the expectation we have developed.

Theorem 4.56. *Let h be a continuous function on a bounded, closed interval $[a, b]$. Think of $[a, b]$ as a sample space Ω with probability measure P given by normalized length of interval: for $[c, d] \subset [a, b]$, $P([c, d]) = (d - c)/(b - a)$. Define the random variable X on Ω by $X(\omega) = h(\omega)$ for $\omega \in \Omega = [a, b]$. Then*

$$EX = \frac{1}{b-a} \int_a^b h(x)\, dx$$

where on the right we have the Riemann integral of h.

Proof. We assume first that $X = h$ is nonnegative. For each positive integer n, partition $[a, b]$ by $a = x_0 < x_1 < \cdots < x_n = b$ where the subintervals have length $\Delta x = x_i - x_{i-1} = (b - a)/2^n$. Choose points $c_i \in [x_{i-1}, x_i]$ so that

$$h(c_i) = \min_{x \in [x_{i-1}, x_i]} h(x), \qquad \text{for } i = 1, \dots, n.$$

Points c_i exist because a continuous functions achieves its minimum on a compact set.

Define the simple random variables

$$X_n(\omega) = h(c_1)\, I_{[x_0,x_1]}(\omega) + \sum_{i=2}^{n} h(c_i)\, I_{(x_{i-1},x_i]}(\omega) \qquad \text{for } a \le \omega \le b.$$

(The first term is separated from the rest simply because it is the only term where the interval contains also its left endpoint.) Because of the choice of dyadic intervals, the sequence $X_n(\omega)$ is nondecreasing.

Then

$$\begin{aligned} \sup_{\omega} |X_n(\omega) - X(\omega)| &= \max_{1 \le i \le n} \sup_{x \in [x_{i-1}, x_i]} |h(c_i) - h(x)| \\ &\le \sup_{|x-y| \le (b-a)/n} |h(x) - h(y)| \to 0 \end{aligned}$$

as $n \to \infty$, by the uniform continuity of h.

We have now verified that $0 \le X_n \nearrow X$ and hence we may take the limit to find the expectation:

$$\begin{aligned} EX &= \lim_{n\to\infty} E[X_n] = \lim_{n\to\infty} \sum_{i=1}^{n} h(c_i)\, P((x_{i-1}, x_i]) = \frac{1}{b-a} \lim_{n\to\infty} \sum_{i=1}^{n} h(c_i)\, \Delta x \\ &= \frac{1}{b-a} \int_a^b h(x)\, dx. \end{aligned}$$

The last equality above is the definition of $\int_a^b h(x)\,dx$ in terms of a limit of Riemann sums.

The case of a general real h comes by decomposition into positive and negative parts. □

Exercises

Exercise 4.1.

Find the expected value of the random variable X in each of the following cases:

(1) $X \sim \mathrm{Exp}(\lambda)$

(2) X is a uniformly chosen element of the set $\{1, 2, 4, \dots, 2^{99}\}$.

Exercise 4.2.

(1) Find all $\alpha > 0$ for which there is a finite constant c_α so that the function $f_\alpha(x) = c_\alpha \frac{1}{1+|x|^\alpha}$ is a PDF.

(2) Suppose that $\alpha > 0$ is a number for which f_α is a PDF. Suppose that X is a random variable with PDF f_α. For which $\alpha > 0$ will $E[X]$ exist?

(3) Suppose that X has a PDF given by f_α and X has a finite expectation. What is $E[X]$?

Hint. Do not try to explicitly compute c_α or the expectation. You need to estimate the integrals.

Exercise 4.3. Let n be a positive integer and p and r two real numbers in $(0, 1)$. Two random variables X and Y are defined on the same sample space. *All* we know about them is that $X \sim \mathrm{Geom}(p)$ and $Y \sim \mathrm{Bin}(n, r)$. For each expectation in parts (a)–(c) below, decide whether it can be calculated with this information. If it can, give its value. If it cannot, explain why.

(a) $E[X + Y]$

(b) $E[X^2 + Y^2]$

(c) $E[\,(X + Y)^2\,]$

Exercise 4.4. Suppose that X has $\mathrm{Gamma}(r, \lambda)$ distribution for some $r, \lambda > 0$. (See (3.41).) Show that $E[X]$ is finite, and find its value.
Hint. Try the $\lambda = 1$ case first using (3.40).

Exercise 4.5. Suppose X is a discrete random variable such that $P(X \in \mathbb{Z}_{\geq 0}) = 1$. Use formula (4.1) to show that $E(X) = \sum_{n=0}^{\infty} P(X > n)$.
Hint. The possible values of X are the non-negative integers, so $E[X] = \sum_{a=0}^{\infty} aP(X = a)$.

Exercise 4.6. Prove Theorem 4.15.

Hint. If $s > 1$ and $X(\omega) > s$, then also $X(\omega)^r > s$. Use (4.12).

Exercise 4.7. Let (X, Y) be jointly absolutely continuous with joint density function f. Assume that expectations EX and EY are finite. Show that, for any real numbers a, b, c,

$$E[aX + bY + c] = aE(X) + bE(Y) + c. \tag{4.51}$$

Exercise 4.8. Suppose $P(X \geq 0) = 1$ and $EX = 0$. Show that then $P(X = 0) = 1$.

Hint. Assume the contrary and use the result of Exercise 2.19 together with the monotonicity of the expectation.

Exercise 4.9. Let $0 < p < 1$ and $X \sim \text{Geom}(p)$. Find the mean and variance of X.

Hint. Power series can be differentiated term by term inside their radius of convergence. For example, $f(x) = \sum_{k=0}^{\infty} x^k$ converges for $|x| < 1$, and formulas $f'(x) = \sum_{k=1}^{\infty} kx^{k-1}$ and $f''(x) = \sum_{k=2}^{\infty} k(k-1)x^{k-2}$ can be used to evaluate the series.

Exercise 4.10. Let X be a $\mathbb{Z}_{\geq 0}$-valued discrete random variable with probability mass function

$$P(X = 0) = \tfrac{4}{5} \quad \text{and} \quad P(X = k) = \tfrac{1}{10} \cdot \left(\tfrac{2}{3}\right)^k \quad \text{for } k \in \{1, 2, 3, \dots\}.$$

(a) Calculate the mean and variance of X.

(b) Identify the distribution of X as that of $I \cdot Y$ for a Bernoulli variable I and an independent random variable Y with a familiar named distribution. Then use this and the results of Exercise 4.9 to give quicker solutions to part (a).

Exercise 4.11. Let $0 < \lambda < \infty$ and $X \sim \text{Exp}(\lambda)$. Find the mean and variance of X.

Exercise 4.12. Recall Example 2.32 from Section 2.1. In units of 1000 dollars, the cost of repairs when the tree falls is uniform in the interval $[1, 5]$ and your insurance has a deductible of 2. Y denotes the cost of the repairs and X the amount you pay, both in units of 1000 dollars. Find the expectations EY and EX.

Hint. Look at Example 4.13.

Exercise 4.13. Let X be uniform on the set $\{-1, 0, 1\}$ and $Y = X^2$. Check that X and Y are uncorrelated but not independent.

Exercise 4.14. Let X and Y have joint density function $f(x, y) = x + y$ on the unit square $\{(x, y) : 0 \leq x, y \leq 1\}$.

(a) Calculate the covariance $\text{Cov}(X, Y)$.

(b) Let $U = X \vee Y$ be the maximum of X and Y. Calculate the mean and variance of U.

Exercise 4.15. I have four different sweaters. Every day I choose one of the four sweaters at random to wear. Let X be the number of different sweaters I wore during a 5-day week. (For example, if my 5-day sweater sequence is $(3, 2, 4, 4, 2)$ then $X = 3$ because that week I wore sweaters 2, 3 and 4.)

(a) Find the mean of X.

(b) Find the variance of X.

Exercise 4.16. I roll a fair die four times. Let X be the number of different outcomes that I saw. (For example, if the die rolls are 5,3,6,6 then $X = 3$ because the different outcomes are 3, 5 and 6.)

(a) Find the mean of X.

(b) Find the variance of X.

Hint. Let I_k be the indicator of the event that the number k appears at least once among the four die rolls. You can safely assume $I_1, I_2, \ldots, I_6$ exchangeable because their probabilities do not care about the particular labeling of the sides of the die.

Exercise 4.17. A cereal company announces a collection of 10 different toys for prizes. Every box of cereal contains two different randomly chosen toys from the series of 10. Let X be the number of different types of toys I have accumulated after buying 4 boxes of cereal.

(a) Find the mean of X.

(b) Find the variance of X.

Exercise 4.18. Suppose that a professor chooses a random student in a class of 40 students (there are 23 girls and 17 boys in the class) to perform a calculation on the board. The professor repeats this procedure 15 times, choosing a new student each time (i.e. no student will go twice). Let X be the total number of boys chosen. Calculate the mean and variance of X.

Exercise 4.19.

(a) Derive the inequality $(a+b)^2 \le 2a^2 + 2b^2$ for real numbers.

(b) Derive the inequality $\operatorname{Var}(X+Y) \le 2\operatorname{Var}(X) + 2\operatorname{Var}(Y)$.

Exercise 4.20. Derive the inequalities

$$\big(\mathrm{SD}(X) - \mathrm{SD}(Y)\big)^2 \le \operatorname{Var}(X+Y) \le \big(\mathrm{SD}(X) + \mathrm{SD}(Y)\big)^2.$$

Exercises for Section 4.3♣.

Exercise 4.21. With $g_n : [0,\infty] \to [0,\infty)$ as in (4.38), show that for each $x \in [0,\infty]$, $g_n(x)$ is a nondecreasing sequence that converges to x.

Exercise 4.22. Show that for any nonnegative random variable X we have the formula

$$EX = \int_0^\infty P(X > s)\, ds.$$

Hint. Prove it first for the truncated variable $X_M = XI_{\{|X| \le M\}}$ as in the proof of Theorem 4.55.

Exercise 4.23. Suppose $X_1, \ldots, X_n$ are independent. Assume that the expectations of all products of distinct X_is are finite. Prove that

$$E\Big[\prod_{i=1}^n X_i\Big] = \prod_{i=1}^n E[X_i].$$

Hint. Start with the case of two nonnegative random variables.

Exercise 4.24 (Dominated convergence theorem)**.** Let X and a sequence X_n be real-valued random variables on $(\Omega, \mathcal{F}, P)$. Assume that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.

(a) Assume that there exists a nonnegative random variable Y on $(\Omega, \mathcal{F}, P)$ that satisfies $EY < \infty$ and also $|X_n(\omega)| \le Y(\omega)$ for all $\omega \in \Omega$ and $n \in \mathbb{Z}_{>0}$. Show that the expectations EX and EX_n are finite and $EX_n \to EX$.
Hint. Let $\underline{X}_n(\omega) = \inf_{k:k \ge n} X_k(\omega)$ and $\bar{X}_n(\omega) = \sup_{k:k \ge n} X_k(\omega)$. Show that $0 \le Y + \underline{X}_n \nearrow Y + X$ and $0 \le Y - \bar{X}_n \nearrow Y - X$.

(b) Show by example that without the assumption in part (a), $EX_n \to EX$ can fail even when $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ and all expectations are finite.

Exercise 4.25. Let h and the sequence h_n be continuous real-valued functions on the compact interval $[a, b]$ and C a constant. Assume that $h_n(x) \to h(x)$ for each $x \in [a, b]$ and $|h_n(x)| \le C$ for each $x \in [a, b]$ and $n \in \mathbb{Z}_{>0}$. Show that

$$\lim_{n\to\infty} \int_a^b h_n(x)\, dx = \int_a^b h(x)\, dx.$$

The three big convergence theorems of integration theory are the monotone convergence theorem, the dominated convergence theorem, and Fatou's lemma. The next exercise rounds off the triple. The particular strength of Fatou's lemma is that it does not assume a limit to begin with.

Exercise 4.26 (Fatou's lemma)**.** Let $\{X_n\}$ be a sequence of $[0, \infty]$-valued random variables on $(\Omega, \mathcal{F}, P)$.

(a) Show that $E\big[\varliminf_{n\to\infty} X_n\big] \le \varliminf_{n\to\infty} E[X_n]$.

(b) Show by an example that strict inequality $<$ can happen in the inequality above.

Exercise 4.27. Prove Corollary 4.40.

Hint. For $0 < M < \infty$, apply Theorem 4.39 to the truncated absolute values $X^M = |X| \wedge M$ and $Y^M = |Y| \wedge M$ to get

$$E[X^M Y^M] \le \sqrt{E[\,|X|^2 \wedge M^2]}\sqrt{E[\,|Y|^2 \wedge M^2]} \le \sqrt{E[X^2]}\sqrt{E[Y^2]}.$$

Let $M \nearrow \infty$.

Chapter 5

Law of large numbers

If we flip a fair coin a large number of times then we expect to see tails 'roughly' half of the time. This seems like a natural statement, but in fact it is a consequence of a theorem called the *law of large numbers* (or LLN for short). The law of large numbers states that if $X_1, X_2, \dots$ are i.i.d. random variables with finite mean $\mu = E[X_1]$ then the average of the first n terms $\frac{1}{n}\sum_{k=1}^n X_k$ converges to μ as $n \to \infty$. Since $\frac{1}{n}\sum_{k=1}^n X_k$ is a random variable, we have to precisely define what it means if a sequence of random variables converges, in fact we will soon see that there are several different formulations.

Returning to the coin flip example, if X_k is the indicator of the event that the kth flip is tails then $\sum_{k=1}^n X_k$ gives the number of tails among the first n flips, and the law of large number states that as $n \to \infty$ the ratio of tails converges to $1/2$ (as we expected).

Here is a quick computation showing why the result of the law of large numbers is 'believable'. Suppose that $X_1, X_2, \dots$ are i.i.d. random variables with finite mean $\mu = E[X_1]$ and finite variance $\sigma^2 = \mathrm{Var}(X_1)$. The expected value of the average of the first n term can be computed using linearity:

$$E[\frac{1}{n}\sum_{k=1}^n X_k] = \frac{1}{n}E[\sum_{k=1}^n X_k] = \frac{1}{n}\sum_{k=1}^n E[X_k] = \frac{1}{n}\cdot n\mu = \mu.$$

We can also compute the variance, using the fact that $X_1, \dots, X_n$ are independent:

$$\mathrm{Var}(\frac{1}{n}\sum_{k=1}^n X_k) = \frac{1}{n^2}\mathrm{Var}(\sum_{k=1}^n X_k) = \frac{1}{n^2}\sum_{k=1}^n \mathrm{Var}(X_k) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{1}{n}\sigma^2.$$

Hence, the average of the first n terms has the same mean as a single term, but the variance is $\frac{1}{n}$ of the variance of a single term. This means that as n gets larger and larger the average of the first n terms gets more and more 'concentrated' around μ. In the first section we develop tools that will allow us to make this statement precise. The further sections show how one prove various versions of the law of large numbers.

5.1. Inequalities of Markov and Chebyshev

Lemma 5.1 (Markov's inequality). *Let X be a $[0,\infty]$-valued random variable and $c > 0$. Then*

$$P(X \ge c) \le \frac{EX}{c}.$$

This inequality is useful only if $c > EX$.

Proof. By considering separately the cases $X(\omega) \ge c$ and $0 \le X(\omega) < c$, verify that

$$X(\omega) \ge cI_{\{X\ge c\}}(\omega) \quad \text{for all } \omega \in \Omega.$$

Taking expectations on both sides gives $EX \ge cP(X \ge c)$. □

Lemma 5.2 (Chebyshev's inequality). *Let X be a random variable with finite mean μ and finite variance σ^2. Then for any real $c > 0$ we have*

$$P(|X-\mu| \ge c) \le \frac{\sigma^2}{c^2}. \tag{5.1}$$

Proof. Since $|X-\mu| \ge 0$ and $c > 0$, the inequality $|X-\mu| \ge c$ is equivalent to the inequality $(X-\mu)^2 \ge c^2$. Apply Markov's inequality to the random variable $(X-\mu)^2$:

$$P(|X-\mu| \ge c) = P((X-\mu)^2 \ge c^2) \le \frac{E[(X-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}.$$

□

The following inequalities are direct consequences of Chebyshev's inequality:

$$P(X \ge \mu + c) \le \frac{\sigma^2}{c^2} \quad \text{and} \quad P(X \le \mu - c) \le \frac{\sigma^2}{c^2} \qquad \text{for } c > 0, \tag{5.2}$$

because the event $\{|X-\mu| \ge c\}$ contains both events $\{X \ge \mu+c\}$ and $\{X \le \mu-c\}$.

The bounds given by Markov's and Chebyshev's inequalities can be very accurate or far from the truth, depending on the case.

Example 5.3. Suppose a nonnegative random variable X has mean 50.

(a) Give an upper bound for the probability $P(X \ge 60)$.

(b) Suppose that we also know $\operatorname{Var}(X) = 25$. How does your bound for $P(X \ge 60)$ change?

(c) Suppose that we also know that X is binomially distributed. Compare the upper bounds from (a) and (b) with he precise value of $P(X \ge 60)$.

In part (a) we can use Markov's inequality to get

$$P(X \ge 60) \le \frac{E[X]}{60} = \frac{50}{60} = \frac{5}{6}.$$

In part (b) we can use Chebyshev's inequality to get

$$P(X \ge 60) = P(X - E[X] \ge 10) \le P(|X - E[X]| \ge 10) \le \frac{\operatorname{Var}(X)}{10^2} = \frac{25}{100} = \frac{1}{4}.$$

This upper bound is better than the one from Markov's inequality in part (a). However, this is not always the case (Exercise 5.1).

(c) If X is binomial, then $np = 50$ and $np(1-p) = 25$ force $p = 1/2$ and $n = 100$. So $X \sim \text{Bin}(100, 1/2)$. The precise probability is

$$P(X \geq 60) = \sum_{k=60}^{100} \binom{100}{k} 2^{-100} \approx 0.0284.$$

△

In the example above the inequalities gave bounds far away from the true probability. Here is an example where Markov's inequality is tight.

Example 5.4. Let $X \sim \text{Ber}(p)$. Then $P(X \geq 1) = p$, and Markov's inequality gives

$$P(X \geq 1) \leq \frac{EX}{1} = p.$$

△

5.2. Weak law of large numbers

Recall the Definition 3.9 of independent and identically distributed (i.i.d.) random variables: the sequence $\{X_k\}$ is i.i.d. if these random variables are independent and they all have the same distribution, that is, each X_k has the same cumulative distribution function F. The law of large numbers asserts that the average of independent and identically distributed random variables converges to the mean. The distinction between weak and strong laws of large numbers has to do with the type of convergence that is explained in Section 5.3.

Theorem 5.5 (Weak law of large numbers, WLLN)**.** *Let $\{X_k\}_{k\geq 1}$ be i.i.d. random variables with finite mean $\mu = E[X_k]$ and finite variance $\sigma^2 = \text{Var}(X_k)$. Let $S_n = X_1 + \cdots + X_n$. Then for any fixed $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} P\big(\,|\tfrac{S_n}{n} - \mu| \geq \varepsilon\big) = 0. \tag{5.3}$$

Proof. Fix the value of $\varepsilon > 0$. To apply Chebyshev's inequality we calculate the mean and variance of S_n/n. By linearity of expectation

$$E\left(\frac{S_n}{n}\right) = E\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] = \mu$$

and by independence

$$\text{Var}\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\text{Var}\left[\sum_{k=1}^{n} X_k\right] = \frac{1}{n^2}\sum_{k=1}^{n}\text{Var}[X_k] = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}. \tag{5.4}$$

Then Chebyshev's inequality gives

$$P\big(\,|\tfrac{S_n}{n} - \mu| \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2}\text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n\varepsilon^2} \underset{n\to\infty}{\to} 0. \tag{5.5}$$

This proves (5.3). □

Theorem 5.5 holds without the assumption of finite variance as long as the expectation $E(X_1) = \mu$ is finite[1].

Calculation (5.4) above quantifies the cancellation that happens when we average independent random contributions: the standard deviation of the sample average S_n/n is $\sigma^2/\sqrt{n}$.

Here is a numerical example that appeals to the WLLN.

Example 5.6. Show that the probability that fair coin flips yield 51% or more tails converges to zero as the number of flips tends to infinity.

Let S_n denote the number of tails in n fair coin flips. Then $E(S_n/n) = 0.5$.

$$\begin{aligned} P(\text{at least } 51\% \text{ tails in } n \text{ flips}) &= P\big(\tfrac{S_n}{n} \ge 0.51\big) = P\big(\tfrac{S_n}{n} - 0.5 \ge 0.01\big) \\ &\le P\big(\big|\tfrac{S_n}{n} - 0.5\big| \ge 0.01\big) \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

The inequality above comes from the monotonicity of probability: the event $\{\frac{S_n}{n} - 0.5 \ge 0.01\}$ lies inside the larger event $\{|\frac{S_n}{n} - 0.5| \ge 0.01\}$. △

5.3. Borel-Cantelli lemma

Definition of convergence in probability and almost surely.

Theorem 5.5 is called the *weak law of large numbers* because it gives convergence of the probabilities, namely that for any $\varepsilon > 0$

$$P\big(|\tfrac{S_n}{n} - \mu| \ge \varepsilon\big) \to 0,$$

but it does not say that the random sequence S_n/n itself converges. Here are the precise definitions of the two types of convergence involved in this distinction.

Definition 5.7. Let $\{X_n\}_{n\ge 1}$ be a sequence of random variables and X another random variable, all defined on the same probability space $(\Omega, \mathcal{F}, P)$.

(a) We say that X_n **converges to X in probability** as $n \to \infty$ if, for each $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \ge \varepsilon) = 0. \tag{5.6}$$

A common abbreviation is $X_n \xrightarrow{P} X$.

(b) We say that X_n **converges to X almost surely** or **with probability one** if there is an event $\Omega_0 \subset \Omega$ such that $P(\Omega_0) = 1$ and for all $\omega \in \Omega_0$, $X_n(\omega) \to X(\omega)$ as $n \to \infty$. Common abbreviations are $X_n \to X$ a.s. and $X_n \to X$ w.p.1. △

The definition of almost sure convergence $X_n \to X$ can be stated succinctly as $P\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$ or even more tersely as $P(X_n \to X) = 1$. Theorem 5.9 below shows that almost sure convergence implies convergence in probability.

With the definition above, WLLN can be stated as follows: if $S_n = X_1 + \cdots + X_n$ is a sum of i.i.d. random variables with common mean $\mu = E(X_k)$ and finite variance, then $S_n/n \xrightarrow{P} \mu$.

[1] Exercise 5.11 sketches a truncation argument for this that relies on some of the more advanced tools developed in Chapter 4

The next example illustrates the difference between convergence in probability and almost sure convergence.

Example 5.8. This is an example where convergence in probability holds but convergence at individual sample points fails. Imagine a robot that shoots hoops. Suppose that the nth shot misses with probability $1/n$, independently of the other shots. Let

$$X_n = \begin{cases} 1 & \text{if the } n\text{th shot succeeds,} \\ 0 & \text{if the } n\text{th shot misses.} \end{cases}$$

So X_n has probability mass function

$$P(X_n = 1) = \frac{n-1}{n} \quad \text{and} \quad P(X_n = 0) = \frac{1}{n}.$$

We claim that $X_n \to 1$ in probability, but the pointwise convergence $X_n(\omega) \to 1$ fails with probability 1. That is,

$$P\{\omega : \lim_{n\to\infty} X_n(\omega) = 1\} = 0. \tag{5.7}$$

Convergence in probability is immediate: for any $0 < \varepsilon \le 1$,

$$P(\,|X_n - 1| \ge \varepsilon) = P(X_n \ne 1) = P(X_n = 0) = \tfrac{1}{n} \to 0 \qquad \text{as } n \to \infty.$$

We turn to the question of convergence at a fixed ω. Using the definition of convergence of real numbers from analysis, since each X_n takes only values 0 and 1, $X_n(\omega) \to 1$ is equivalent to the statement that

$$\exists m \ge 1 \text{ such that } X_n(\omega) = 1 \text{ for all } n \ge m. \tag{5.8}$$

We show that the event in (5.8) happens with probability zero. First we consider a fixed m:

$$\begin{aligned} P(X_n = 1 \;\forall n \ge m) &= \lim_{\ell\to\infty} P(X_n = 1 \text{ for all } n \in \{m, m+1, \dots, \ell\}) \\ &= \lim_{\ell\to\infty} \prod_{n=m}^{\ell} \frac{n-1}{n} = \lim_{\ell\to\infty} \Big(\frac{m-1}{m} \cdot \frac{m}{m+1} \cdot \frac{m+1}{m+2} \cdots \frac{\ell-1}{\ell} \Big) \\ &= \lim_{\ell\to\infty} \frac{m-1}{\ell} = 0. \end{aligned}$$

Then by subadditivity for the union:

$$\begin{aligned} &P(\,\exists m \ge 1 \text{ such that } X_n(\omega) = 1 \text{ for all } n \ge m\,) \\ &\qquad \le \sum_{m=1}^{\infty} P(X_n = 1 \;\forall n \ge m) = \sum_{m=1}^{\infty} 0 = 0. \end{aligned} \tag{5.9}$$

Thus (5.8) happens with probability zero, which is the same as saying that (5.7) holds.

One way to grasp the message of this example is to think about simulating this process on a computer. Missed shots become less and less likely as time passes, but no matter how long you have waited, there is always another missed shot in the future. This is exactly what (5.9) says: with probability one, no matter how large an m we take, there is always an $n \ge m$ such that the nth shot misses. Hence in fact there are infinitely many missed shots. △

Infinitely often happening events.

We develop concepts for describing the event that a sequence of random variables converges. Recall the definition of a limit of real numbers: $x_n \to x$ if the following holds:

$$\forall k \in \mathbb{Z}_{>0} \; \exists N \in \mathbb{Z}_{>0} \text{ such that for all } n \geq N, \; |x_n - x| < \frac{1}{k}.$$

We replaced the usual real $\varepsilon > 0$ with $1/k$ for positive integers k because in probability it is convenient that $1/k$ has only countably many values. The statement above can be expressed in words as follows: $x_n \to x$ if for every $k \in \mathbb{Z}_{>0}$, $|x_n - x| < \frac{1}{k}$ *for all but finitely many n.*

The complementary statement is that $x_n \not\to x$ if the following holds:

$$\exists k \in \mathbb{Z}_{>0} \text{ such that } \forall N \text{ there exists } n \geq N \text{ such that } |x_n - x| \geq \frac{1}{k}.$$

In other words: $x_n \to x$ fails if for some $k \in \mathbb{Z}_{>0}$, $|x_n - x| \geq 1/k$ happens *for infinitely many* n, or *infinitely often* (if n is understood as the index that repeats).

Adapt this to random variables. For a fixed sample point $\omega \in \Omega$ the statement is the following: $\lim\limits_{n\to\infty} X_n(\omega) = X(\omega)$ fails if for some $k \in \mathbb{Z}_{>0}$, $|X_n(\omega) - X(\omega)| \geq 1/k$ for infinitely many n. In terms of events:

$$\begin{aligned} &\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega) \text{ fails}\} \\ &\qquad = \bigcup_{k\in\mathbb{Z}_{>0}} \{\omega : |X_n(\omega) - X(\omega)| \geq 1/k \text{ for infinitely many } n\}. \end{aligned} \tag{5.10}$$

This example indicates that we need to understand how to say "happens for infinitely many n" in terms of set operations.

Let $\{A_n\}_{n\geq 1}$ be a sequence of events.

$$\begin{aligned} &\{\omega : \omega \in A_n \text{ for infinitely many } n\} \\ &\qquad = \{\omega : \forall m \geq 1 \; \exists n \geq m \text{ such that } \omega \in A_n \;\} \\ &\qquad = \bigcap_{m=1}^{\infty} \{\omega : \exists n \geq m \text{ such that } \omega \in A_n \;\} = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n. \end{aligned} \tag{5.11}$$

This is sometimes also written as

$$\overline{\lim}\, A_n = \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n$$

and called the *limit superior* of A_n. Then we have the identity

$$I_{\overline{\lim}\, A_n}(\omega) = \overline{\lim_{n\to\infty}}\, I_{A_n}(\omega).$$

The complementary event is

$$\begin{aligned} &\{\omega : \omega \in A_n \text{ for only finitely many } n\} \\ &\qquad = \{\omega : \exists m \geq 1 \text{ such that } \forall n \geq m, \; \omega \in A_n^c \;\} = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c. \end{aligned}$$

Note also how de Morgan's laws give the same formula for the complement:

$$\{\omega : \omega \in A_n \text{ for infinitely many } n\}^c = \Bigl(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \Bigr)^c$$
$$= \bigcup_{m=1}^{\infty} \Bigl(\bigcup_{n=m}^{\infty} A_n \Bigr)^c = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n^c.$$

The operation above is sometimes also written as the limit inferior

$$\underline{\lim}\, B_n = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} B_n$$

and expressed in words as the set of ω that lie in all but finitely many B_n. The identity $I_{\underline{\lim}\, B_n}(\omega) = \underline{\lim}_{n\to\infty} I_{B_n}(\omega)$ is valid.

Now we can complete (5.10) into a set expression for the failure of convergence.

$$\begin{aligned} &\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega) \text{ fails}\} \\ &\quad = \bigcup_{k\in\mathbb{Z}_{>0}} \{\omega : |X_n(\omega) - X(\omega)| \ge 1/k \text{ for infinitely many } n\} \\ &\quad = \bigcup_{k\in\mathbb{Z}_{>0}} \bigcap_{m\in\mathbb{Z}_{>0}} \bigcup_{n:\, n\ge m} \{\omega : |X_n(\omega) - X(\omega)| \ge 1/k\}. \end{aligned} \tag{5.12}$$

With the ideas from above, we can check that almost sure convergence implies convergence in probability.

Theorem 5.9. *Suppose $X_n \to X$ almost surely. Then also $X_n \to X$ in probability.*

Proof. Let $\varepsilon > 0$. Below we use monotonicity of events and then (5.10).

$$\begin{aligned} \lim_{m\to\infty} P(\,|X_m - X| \ge \varepsilon) &\le \lim_{m\to\infty} P\Bigl(\bigcup_{n=m}^{\infty} \{|X_n - X| \ge \varepsilon\} \Bigr) \\ &= P\Bigl(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} \{|X_n - X| \ge \varepsilon\} \Bigr) \\ &= P(\,|X_n - X| \ge \varepsilon \text{ happens for infinitely many } n) \\ &\le P(X_n \to X \text{ fails}) = 0. \end{aligned}$$

The assumption of almost sure convergence $X_n \to X$ gave the last equality. We have established that $P(\,|X_n - X| \ge \varepsilon) \to 0$ for all $\varepsilon > 0$, which is the definition of $X_n \overset{P}{\to} X$. □

Borel-Cantelli lemma.

Lemma 5.10 (Borel-Cantelli lemma)**.** *Let $\{A_n\}_{n\ge1}$ be a sequence of events, all defined on the same sample space. Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then*

$$P\{\omega : \omega \in A_n \textit{ for infinitely many } n\} = 0.$$

Proof. We give two proofs. The second one uses results from Section 4.3.

Proof 1. Recall that if a series $\sum_{k=1}^{\infty} x_k$ converges, then $\lim_{n\to\infty} \sum_{k=n}^{\infty} x_k = 0$. Now from (5.11):

$$P\{A_n \text{ happens for infinitely many } n\} = P\biggl(\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty} A_n\biggr) = \lim_{m\to\infty} P\biggl(\bigcup_{n=m}^{\infty} A_n\biggr)$$
$$\le \lim_{m\to\infty} \sum_{n=m}^{\infty} P(A_n) = 0.$$

Above we used the monotonicity of the sequence of events $\bigcup_{n=m}^{\infty} A_n$ as m increases and then that the probability of a finite or countable union is always bounded by the sum of probabilities.

Proof 2. Let $N(\omega) = \sum_{n=1}^{\infty} I_{A_n}(\omega)$ count how many of the events A_n occur. N is a nonnegative random variable with values in $\mathbb{Z}_{\ge 0} \cup \{\infty\}$. Its expectation is well-defined.

$$E[N] = E\biggl[\sum_{n=1}^{\infty} I_{A_n}\biggr] = E\biggl[\lim_{m\to\infty}\sum_{n=1}^{m} I_{A_n}\biggr] = \lim_{m\to\infty} E\biggl[\sum_{n=1}^{m} I_{A_n}\biggr]$$
$$= \lim_{m\to\infty}\sum_{n=1}^{m} E[I_{A_n}] = \lim_{m\to\infty}\sum_{n=1}^{m} P(A_n) = \sum_{n=1}^{\infty} P(A_n) < \infty.$$

The third equality above used the monotone convergence theorem (Theorem 4.54). We conclude that $E[N]$ is finite. This implies that N is finite with probability one (Theorem 4.46). Thus with probability one only finitely many events A_n occur. The complementary event $\{\omega : A_n \text{ happens for infinitely many } n\}$ must have probability zero. □

Example 5.11. In Example 5.8 we found that when independent variables X_n satisfy

$$P(X_n = 1) = 1 - \tfrac{1}{n} \quad \text{and} \quad P(X_n = 0) = \tfrac{1}{n}$$

then $X_n \overset{P}{\to} X$ but $X_n(\omega) \to X(\omega)$ fails with probability one. Change the rules to

$$P(X_n = 1) = 1 - \tfrac{1}{n^2} \quad \text{and} \quad P(X_n = 0) = \tfrac{1}{n^2}.$$

Now

$$\sum_{n=1}^{\infty} P(X_n = 0) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

and so by the Borel-Cantelli lemma,

$$P\{\omega : X_n(\omega) = 0 \text{ only finitely many times }\} = 1.$$

To put this in other words, we can define a random variable

$$K(\omega) = \sup\{n : X_n(\omega) = 0\}$$

as the last shot that the robot misses (or infinite if there is always a next miss), and then we know that

$$P\{\omega : K(\omega) < \infty\} = P\{\omega : X_n(\omega) = 0 \text{ only finitely many times}\} = 1.$$

In particular, $n > K(\omega)$ implies that $X_n(\omega) = 1$ so $X_n \to 1$ w.p.1.

However, *we can never know for sure if we have seen the last miss.* We can only calculate probabilities. The probability that after m shots we have seen all the misses is

$$\begin{aligned}
P(K \le m) &= P(X_n = 1 \text{ for all } n \ge m+1) = \lim_{\ell\to\infty} \prod_{n=m+1}^{\ell} \frac{n^2-1}{n^2} \\
&= \lim_{\ell\to\infty} \prod_{n=m+1}^{\ell} \frac{(n-1)(n+1)}{n^2} = \lim_{\ell\to\infty} \frac{\prod_{n=m}^{\ell-1} n \cdot \prod_{n=m+2}^{\ell+1} n}{\prod_{n=m+1}^{\ell} n \cdot \prod_{n=m+1}^{\ell} n} \\
&= \lim_{\ell\to\infty} \frac{m(\ell+1)}{\ell(m+1)} = \frac{m}{m+1}.
\end{aligned}$$

△

Almost sure convergence from the Borel-Cantelli lemma.

The next theorem encapsulates the most common strategy for proving almost sure convergence with the Borel-Cantelli lemma.

Theorem 5.12. *Let $\{X_n\}_{n\ge1}$ and X be random variables on $(\Omega, \mathcal{F}, P)$. Suppose that for all $\varepsilon > 0$,*

$$\sum_{n=1}^{\infty} P(\,|X_n - X| \ge \varepsilon) < \infty. \tag{5.13}$$

Then $X_n \to X$ almost surely.

Proof. For $k \in \mathbb{Z}_{>0}$ define the event

$$B_k = \{\,|X_n - X| \ge \tfrac{1}{k} \text{ happens for only finitely many } n\}.$$

Assumption (5.13) with $\varepsilon = 1/k$ gives

$$\sum_{n=1}^{\infty} P(\,|X_n - X| \ge \tfrac{1}{k}) < \infty$$

and then the Borel-Cantelli lemma gives the conclusion

$$P(\,|X_n - X| \ge \tfrac{1}{k} \text{ happens for infinitely many } n) = 0.$$

This says that $P(B_k) = 1$.

Let $B = \bigcap_{k\ge1} B_k$. Then $P(B) = 1$. To complete the proof, we observe that $X_n \to X$ on the event B. Let $\omega \in B$. Then $\omega \in B_k$ for each positive integer k. By the definition of B_k, there exists a finite index m_k (that depends on ω) such that $|X_n(\omega) - X(\omega)| < \frac{1}{k}$ for $n \ge m_k$. Since this is true for each k, this says precisely that $X_n(\omega) \to X(\omega)$. □

Remark 5.13. Comparison of condition (5.13) with definition (5.6) of convergence in probability illuminates the difference between the two: if $P(\,|X_n - X| \ge \varepsilon)$ converges to zero fast enough to be summable (as in (5.13)) then convergence in probability can be upgraded to almost sure convergence.

However, let us also note that (5.13) is sufficient for almost sure convergence but not necessary (Exercise 5.9 sketches an example). △

5.4. Strong law of large numbers

Theorem 5.14 (Strong law of large numbers, SLLN). *Let $\{X_k\}_{k\ge 1}$ be i.i.d. random variables with finite mean $\mu = E[X_k]$. Let $S_n = X_1 + \cdots + X_n$. Then*

$$P\Big\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = \mu\Big\} = 1.$$

In other words, $S_n/n \to \mu$ almost surely.

Proof. We give a proof under the assumption of a finite fourth moment: $E(X_k^4) < \infty$. This implies also that $E[|X|^s] < \infty$ for all $1 \le s \le 4$ (Theorem 4.15 proved in Exercise 4.6). We give again two proofs. Both proofs use an estimation of the fourth moment of $S_n - n\mu$.

Proof 1. The proof idea is to use Markov's inequality to show that, for each fixed $\varepsilon > 0$,

$$\sum_{n=1}^{\infty} P\big\{\big|\tfrac{S_n}{n} - \mu\big| \ge \varepsilon\big\} < \infty. \tag{5.14}$$

Then Theorem 5.12 finishes the proof.

To simplify notation, let $\bar X_k = X_k - \mu$ and $\bar S_n = \sum_{k=1}^n \bar X_k = S_n - n\mu$. The moment estimate is still valid for $\bar X_k$: for $a, b \ge 0$,

$$(a+b)^4 \le (2(a\vee b))^4 \le 2^4(a^4 \vee b^4) \le 16(a^4 + b^4),$$

and consequently

$$E[(\bar X_k)^4\,] = E[(X_k-\mu)^4\,] \le E[(\,|X_k| + |\mu|\,)^4\,] \le 16E[X_k^4\,] + 16\mu^4 < \infty.$$

Consequently we have again $E[|\bar X_k|^s] < \infty$ for all $1 \le s \le 4$, so all moments of $\bar X_k$ up to the fourth are finite.

Apply Markov's inequality with a fourth moment:

$$\begin{aligned} P\big\{\big|\tfrac{S_n}{n} - \mu\big| \ge \varepsilon\big\} &= P\big\{\big|\tfrac{\bar S_n}{n}\big| \ge \varepsilon\big\} = P\big\{\big|\tfrac{\bar S_n}{n}\big|^4 \ge \varepsilon^4\big\} \\ &\le \frac{1}{\varepsilon^4} E\big(\,\big|\tfrac{\bar S_n}{n}\big|^4\,\big) = \frac{1}{\varepsilon^4 n^4} E[(\bar S_n)^4\,]. \end{aligned}$$

Next we expand the moment.

$$\begin{aligned} E[(\bar S_n)^4\,] &= E\bigg[\bigg(\sum_{k=1}^n \bar X_k\bigg)^4\,\bigg] = E\bigg[\sum_{1\le i,j,k,\ell\le n} \bar X_i \bar X_j \bar X_k \bar X_\ell\bigg] \\ &= \sum_{1\le i,j,k,\ell\le n} E\big[\bar X_i \bar X_j \bar X_k \bar X_\ell\big] \\ &= \sum_{i=1}^n E\big[\bar X_i^4\big] + 6 \sum_{1\le i<j\le n} E\big[\bar X_i^2 \bar X_j^2\big] \\ &= nE[\bar X_1^4] + 3n(n-1)E[\bar X_1^2]E[\bar X_2^2]. \end{aligned}$$

The key to the calculation above is that if index i is different from j, k, ℓ, then by independence,

$$E\big[\bar X_i \bar X_j \bar X_k \bar X_\ell\big] = E\big[\bar X_i\big]E\big[\bar X_j \bar X_k \bar X_\ell\big] = 0 \cdot E\big[\bar X_j \bar X_k \bar X_\ell\big] = 0.$$

So the only types of terms that survive are $E[\bar{X}_i^4]$ and $E[\bar{X}_i^2\bar{X}_j^2]$ for $i<j$. The factor $6=\binom{4}{2}$ in front of $E[\bar{X}_i^2\bar{X}_j^2]$ comes as follows: for each pair (i,j) of indices $i<j$, we pick two out of the four $\bar{X}$-factors to be $\bar{X}_i$ and the other two to be $\bar{X}_j$.

For our present purpose we do not care about the exact values of the moments $E[\bar{X}_1^4]$ and $E[\bar{X}_1^2]$. Hence we introduce a constant C that bounds their values and then we can write

$$E[(\bar{S}_n)^4\,]\le Cn+Cn(n-1)=Cn^2. \tag{5.15}$$

Now combine the steps from above.

$$\sum_{n=1}^{\infty}P\{\bigl|\tfrac{S_n}{n}-\mu\bigr|\ge\varepsilon\}\le\sum_{n=1}^{\infty}\frac{1}{\varepsilon^4n^4}E[(\bar{S}_n)^4\,]\le\sum_{n=1}^{\infty}\frac{C}{\varepsilon^4n^2}<\infty.$$

(5.14) has been verified and now Theorem 5.12 implies that $\frac{S_n}{n}-\mu\to0$ almost surely. The SLLN has been proved under the fourth moment assumption.

We can finish the argument in a different way that avoids the use of Theorem 5.12. Instead of a fixed $\varepsilon>0$, we take ε varying with n: $\varepsilon_n=n^{-1/8}$. Then we have $\varepsilon_n\to0$ while still making the series converge: from the bound above,

$$\sum_{n=1}^{\infty}P\{\bigl|\tfrac{S_n}{n}-\mu\bigr|\ge n^{-1/8}\}\le\sum_{n=1}^{\infty}\frac{C}{(n^{-1/8})^4n^2}=\sum_{n=1}^{\infty}\frac{C}{n^{3/2}}<\infty.$$

The Borel-Cantelli lemma implies that, with probability one, the event $\{\omega:\bigl|\frac{S_n(\omega)}{n}-\mu\bigr|\ge n^{-1/8}\}$ happens for only finitely many n. Consequently, there exists a random variable $N(\omega)$ that is finite with probability one and such that for $n>N(\omega)$ we have

$$\bigl|\tfrac{S_n(\omega)}{n}-\mu\bigr|<n^{-1/8}.$$

This last inequality implies that $\frac{S_n(\omega)}{n}\to\mu$ as $n\to\infty$.

Proof 2. This proof is in spirit similar to the second proof of the Borel-Cantelli lemma. This is now quick because we can use estimate (5.15) from above.

$$E\biggl[\,\sum_{n=1}^{\infty}\Bigl(\frac{\bar{S}_n}{n}\Bigr)^4\biggr]=\sum_{n=1}^{\infty}\frac{E[(\bar{S}_n)^4\,]}{n^4}\le\sum_{n=1}^{\infty}\frac{C}{n^2}<\infty.$$

The first equality above used the monotone convergence theorem (Theorem 4.54) to take the sum outside the expectation.

The series $\sum_{n=1}^{\infty}(\bar{S}_n/n)^4$ is a nonnegative random variable. Since it has a finite expectation, it must be finite with probability one. In other words, the event where the series converges has probability one:

$$P\biggl\{\omega:\sum_{n=1}^{\infty}\Bigl(\frac{\bar{S}_n(\omega)}{n}\Bigr)^4\ \text{converges}\ \biggr\}=1. \tag{5.16}$$

Suppose ω lies in the event above. Then, since terms of a convergent series themselves converge to zero, we must have $\lim_{n\to\infty}(\bar{S}_n(\omega)/n)^4=0$. This is equivalent to $\lim_{n\to\infty}\bar{S}_n(\omega)/n=0$, which is the same as saying that $\lim_{n\to\infty}S_n(\omega)/n=\mu$. We have shown that $\lim_{n\to\infty}S_n/n=\mu$ holds on the probability one event in (5.16) and thereby proved that this convergence happens with probability one. □

Example 5.15. To understand that we cannot expect convergence $\frac{1}{n}S_n(\omega) \to \mu$ for all $\omega \in \Omega$, but only on an event of probability one, it is illuminating to return to independent trials. The sample space is the space of $\{0,1\}$-valued sequences

$$\Omega = \{\omega = (s_i)_{i \in \mathbb{Z}_{>0}} : \text{each } s_i \text{ equals } 0 \text{ or } 1\}$$

and the random variables are the coordinate functions: $X_k(\omega) = s_k$. For each success probability $p \in [0,1]$, we put a probability measure P_p on Ω under which the variables $\{X_k\}$ are i.i.d. Ber(p) variables: for any n-tuple $(t_1, \dots, t_n) \in \{0,1\}^n$:

$$P_p\{X_1 = t_1,\, X_2 = t_2, \dots, X_n = t_n\} = p^{\sum_{i=1}^n t_i}(1-p)^{\sum_{i=1}^n (1-t_i)}.$$

Let

$$A_p = \Big\{\omega : \lim_{n\to\infty} \frac{S_n(\omega)}{n} = p\Big\}$$

denote the set of sequences ω whose average converges to p. Then the strong law of large numbers says precisely that $P_p(A_p) = 1$. For different values of p, the events A_p are disjoint. Hence each probability measure P_p puts all its probability on a particular event A_p disjoint from all the others.

The sets A_p do not account for all of Ω because there are sequences ω whose averages do not converge. For example, define $\omega = (s_i)_{i\in\mathbb{Z}_{>0}}$ by

$$s_i = \begin{cases} 1, & \text{if } 4^{2m} < i \le 4^{2m+1} \text{ for some } m \in \mathbb{Z}_{\ge 0} \\ 0, & \text{if } 4^{2m+1} < i \le 4^{2(m+1)} \text{ for some } m \in \mathbb{Z}_{\ge 0}. \end{cases}$$

Then $4^{-2m-1}S_{4^{2m+1}}(\omega) \ge \frac{3}{4}$ while $4^{-2(m+1)}S_{4^{2(m+1)}}(\omega) \le \frac{1}{4}$ for each m, and so $n^{-1}S_n(\omega)$ cannot converge. △

Fluctuations: beyond the law of large numbers.

Recall the estimate that gave us the WLLN:

$$P\Big(\Big|\frac{S_n}{n} - \mu\Big| \ge \varepsilon\Big) \le \frac{1}{\varepsilon^2}\operatorname{Var}\Big(\frac{S_n}{n}\Big) = \frac{\sigma^2}{n\varepsilon^2}. \tag{5.17}$$

This tells us that in the $n \to \infty$ limit, all the probability distribution of S_n/n concentrates in the interval $(\mu - \varepsilon, \mu + \varepsilon)$, and that this is true for each $\varepsilon > 0$. Imagine focusing a microscope on the real line around μ. Is the statement still true for a smaller interval of size $(\mu - cn^{-\alpha}, \mu + cn^{-\alpha})$ for constants $c > 0$ and $\alpha > 0$? Repeating the estimate with $\varepsilon = cn^{-\alpha}$ gives

$$P\Big(\Big|\frac{S_n}{n} - \mu\Big| \ge \frac{c}{n^\alpha}\Big) \le \frac{n^{2\alpha}}{c^2}\operatorname{Var}\Big(\frac{S_n}{n}\Big) = \frac{n^{2\alpha}}{c^2}\cdot\frac{\sigma^2}{n} = \frac{\sigma^2}{c^2}n^{2\alpha-1}. \tag{5.18}$$

This vanishes still as $n \to \infty$ no matter how small c is, as long as $0 < \alpha < 1/2$. But at $\alpha = 1/2$ the last bound does not converge to zero and the proof fails.

In other words, when we shrink the interval down to $(\mu - cn^{-1/2}, \mu + cn^{-1/2})$, we cannot claim that all the probability concentrates $(\mu - cn^{-1/2}, \mu + cn^{-1/2})$. Are we seeing merely the failure of a particular proof method, or something genuine? Another key piece of information is that the standard deviation of S_n/n is $\sigma n^{-1/2}$. This suggests to us that S_n/n has stochastic fluctuations of size $n^{-1/2}$. Indeed, on the scale $n^{-1/2}$ we will see a different phenomenon: the stochastic fluctuations do not concentrate in a small interval in the limit, but instead acquire a definite shape of the Gaussian type. This is the content of the *central limit theorem*, one of the main results of probability theory.

Exercises

Exercise 5.1. Let X be a nonnegative random variable with finite, positive mean μ and variance σ^2. Show that for some $c > \mu$, Markov's inequality gives a tighter upper bound for the probability $P(X \geq c)$ than Chebyshev's inequality.

Exercise 5.2. When the cashier comes to work to begin his shift, there are k_0 customers in line waiting to be served (k_0 is a nonrandom integer). Then new customers arrive at random. The times between the arrivals of new customers are i.i.d. random variables Y_i with mean $E(Y_i) = 2$ minutes. (That is, the arrival time of the nth new customer is $S_n = \sum_{i=1}^{n} Y_i$.) The cashier takes exactly one minute to process each customer. When there is nobody in line to be served, the cashier can rest. Show that with probability one, eventually the cashier will get a rest.

Hint. What must be true at time S_n when the nth new customer arrives, *if* the cashier has had no rest by that time? You can solve this either with the strong law or the weak law.

Exercise 5.3 (Weak law of large numbers for sampling without replacement)**.** An urn has N_r red balls and N_g green balls, with a total of $N = N_r + N_g$ balls. Let $1 \leq n \leq N$. Sample n balls *without replacement.* Let S_n be the number of red balls in the sample. (It would be more accurate to write S_{n,N_r,N_g} so that all the parameters of the experiment are included, but for simplicity we write S_n.) Show that the following estimate holds for all choices of $N \geq n \geq 1$, $N_r \geq 1$, $N_g \geq 1$, and $\varepsilon > 0$:

$$P\left\{ \left| \frac{S_n}{n} - \frac{N_r}{N} \right| \geq \varepsilon \right\} \leq \frac{1}{n\varepsilon^2}.$$

Exercise 5.4. Fix $0 < p < 1$. Let $\{X_n\}_{n\geq 1}$ be a sequence of independent Ber(p) random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. Let A be the set of ω such that the sequence $X_n(\omega)$ converges as $n \to \infty$. In other words,

$$A = \{\omega : \lim_{n\to\infty} X_n(\omega) \text{ exists}\}.$$

Find the probability $P(A)$.

Exercise 5.5. Let $\{Y_n\}_{n\geq 1}$ be a sequence of random variables such that Y_n has exponential distribution with parameter n (that is, the mean of Y_n is $1/n$). Does the sequence Y_n converge to a limit almost surely? Prove or disprove.

Exercise 5.6. Let $X_1, X_2, X_3, \ldots$ be i.i.d. Unif$(0, 1)$ random variables. Let $M_n = \max(X_1, \ldots, X_n)$ be the maximum of the first n random numbers.

(a) Show that $M_n \to 1$ in probability.

(b) Show that $M_n \to 1$ almost surely.

Exercise 5.7. Let $X_1, X_2, X_3, \dots$ be a sequence of random variables that all have mean 0 and variance 1, and assume that they are negatively correlated, that is, $\mathrm{Cov}(X_i, X_j) \le 0$ for all pairs $i \ne j$. Let $S_n = X_1 + \cdots + X_n$. Show that $n^{-3/4} S_n$ converges to zero in probability.

Exercise 5.8. Let X be a nonnegative random variable and assume $E(X^4) < \infty$. Use the Cauchy-Schwarz inequality in the form given in Corollary 4.40 to show that $E(X)$, $E(X^2)$ and $E(X^3)$ are finite.

Exercise 5.9. Let K be a positive integer valued random variable on $(\Omega, \mathcal{F}, P)$. Define a sequence $\{X_n\}_{n\ge1}$ of $\{0,1\}$-valued random variables on $(\Omega, \mathcal{F}, P)$ by

$$X_n(\omega) = \begin{cases} 0, & n \le K(\omega) \\ 1, & n > K(\omega). \end{cases}$$

Show that by a suitable choice of the distribution of K, $\sum_{n=1}^{\infty} P(X_n \ne 1) = \infty$. In other words, the Borel-Cantelli lemma cannot detect the almost sure convergence $X_n \to 1$.

Exercise 5.10. Let $\{X_n\}_{n\ge1}$ and X be random variables on $(\Omega, \mathcal{F}, P)$. Let $1 \le p < \infty$. We say that $X_n \to X$ *in pth moment*, or *in* L^p, if

$$\lim_{n\to\infty} E\big[\,|X_n - X|^p\,\big] = 0. \tag{5.19}$$

Show that this implies $X_n \xrightarrow{P} X$.

Challenging problems.

Exercise 5.11. Prove the following version of the weak law of large numbers that assumes only a finite mean instead of a finite second moment.

Theorem 5.16. *Let $\{X_k\}_{k\ge1}$ be i.i.d. random variables with finite mean $\mu = E[X_k]$. Let $S_n = X_1 + \cdots + X_n$. Then for any fixed $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\big(\,|\tfrac{S_n}{n} - \mu| \ge \varepsilon\big) = 0. \tag{5.20}$$

Hint. For $0 < M < \infty$, define truncated variables $X_k^M(\omega) = X_k(\omega) I_{\{|X_k| \le M\}}(\omega)$ with mean $\mu^M = E[X_k^M]$ and $S_n^M = X_1^M + \cdots + X_n^M$. Given $\varepsilon, \delta \in (0,1)$, use the dominated convergence theorem (Exercise 4.24) to show that by choosing M large enough,

$$|\mu - \mu^M| \le E|X_1 - X_1^M| \le \varepsilon\delta/4.$$

By the triangle inequality,

$$P\big(\,|S_n - n\mu| \ge n\varepsilon\big) \le P\big(\,|S_n - S_n^M| \ge n\varepsilon/4\big) + P\big(\,|S_n^M - n\mu^M| \ge n\varepsilon/4\big).$$

On the right-hand side above, apply Markov's inequality to the first probability and Chebyshev's inequality to the second probability. Let $n \to \infty$ while keeping M fixed.

Chapter 6

Limits in distribution

6.1. Convergence in distribution

The definition of a limit in distribution is given below. In contrast with almost sure convergence and convergence in probability, for convergence in distribution the random variables do *not* have to be defined on the same sample space.

Definition 6.1. Suppose that for each positive integer n, X_n is a random variable with cumulative distribution function F_n. Let X be a random variable with cumulative distribution function F. Then X_n **converges to** X **in distribution** if $F_n(x) \to F(x)$ as $n \to \infty$ at each point x where F is continuous.

Convergence in distribution is also called *convergence in law* and *weak convergence*. Common abbreviations are $X_n \overset{d}{\to} X$ and $X_n \Rightarrow X$. Since the convergence is defined in terms of the cumulative distribution functions, it is also expressed as $F_n \overset{d}{\to} F$ and $F_n \Rightarrow F$.

The limit $F_n \overset{d}{\to} F$ determines the values $F(x)$ only at continuity points of F. Such points are dense in $\mathbb{R}$, so by right-continuity of F, they are sufficient for unique determination of the entire cumulative distribution function F (Exercise 2.17). The definition requires convergence $F_n(x) \to F(x)$ only at continuity points of F to accommodate cases where F is not continuous. This is illustrated by the next basic example.

Example 6.2. Let a_n and a be real numbers that satisfy $a_n \neq a$ but $a_n \to a$. Let X_n with cumulative distribution function F_n and X with cumulative distribution function F be degenerate random variables that satisfy $P(X_n = a_n) = 1$ and $P(X = a) = 1$. Their cumulative distribution functions are

$$F_n(x) = \begin{cases} 0, & x < a_n \\ 1, & x \geq a_n \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0, & x < a \\ 1, & x \geq a. \end{cases}$$

Consider any $x \neq a$. For large enough n, both a and a_n are strictly on one side of x, and then $F_n(x) = F(x)$. Thus $F_n(x) \to F(x)$ at all $x \neq a$, which is exactly at all continuity points x of F.

For the limit of $F_n(a)$, consider two cases.

(i) If $a > a_n$ for all n, then $F(a) = 1 = F_n(a)$ for all n. Consequently $F_n(a) \to F(a)$, and together with the above, $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$.

(ii) If $a < a_n$ for all n, then $F(a) = 1$ while $F_n(a) = 0$ for all n. Consequently F_n does not converge to F at the jump point a of F.

Thus if convergence in distribution required $F_n(x) \to F(x)$ at *every* x instead of only at continuity points of F, we would quite artificially accept the limit $F_n \overset{d}{\to} F$ if a_n approaches a from the left, but not if a_n approaches a from the right. Under Definition 6.1 we have the natural outcome: $a_n \to a$ implies the weak convergence of the degenerate random variables, no matter how a_n approaches a. △

In the next two examples we cast the binomial-to-Poisson and geometric-to-exponential limits in the framework of Definition 6.1.

Example 6.3. Theorem 3.31 showed the following limit. With $0 < \lambda < \infty$ and $n \in \mathbb{Z}_{>0}$ large enough so that $\lambda/n < 1$, let $S_n \sim \text{Bin}(n, \lambda/n)$ and $Z \sim \text{Poisson}(\lambda)$. Then $P(S_n = k) \to P(Z = k)$ as $n \to \infty$, for each $k \in \mathbb{Z}_{\geq 0}$. From this we get for $t \geq 0$,

$$F_{S_n}(t) = P(S_n \leq t) = \sum_{k=1}^{\lfloor t \rfloor} P(S_n = k) \underset{n\to\infty}{\longrightarrow} \sum_{k=1}^{\lfloor t \rfloor} P(Z = k) = F_Z(t).$$

The limit of the sum above is legitimate because the sum has only finitely many terms. For $t < 0$, $F_{S_n}(t) = F_Z(t) = 0$. Thus we have $F_{S_n}(t) \to F_Z(t)$ for all real t, and thereby the distributional limit $S_n \Rightarrow Z$. △

Example 6.4. Theorem 3.37 gave the following result. With $0 < \lambda < \infty$ and $n \in \mathbb{Z}_{>0}$ large enough so that $\lambda/n < 1$, let T_n satisfy $nT_n \sim \text{Geom}(\lambda/n)$ and $Y \sim \text{Exp}(\lambda)$. Then $P(T_n > t) \to P(Y > t)$ as $n \to \infty$, for all real $t \geq 0$. This gives $F_{T_n}(t) \to F_Y(t)$ for all $t \in \mathbb{R}$, since the cumulative distribution functions vanish for $t < 0$. Thus we have the weak limit $T_n \overset{d}{\to} Y$. △

As the term weak convergence suggests, this type of convergence is the weakest among the ones we have encountered.

Theorem 6.5. *Suppose $X_n \to X$ in probability. Then also $X_n \to X$ in distribution.*

Proof. Let F_n be the distribution function of X_n and F that of X. Let F be continuous at x. We need to show that $F_n(x) \to F(x)$ as $n \to \infty$, equivalently, that $P(X_n \leq x) \to P(X \leq x)$.

We organize the proof to show both

$$\overline{\lim_{n\to\infty}}\, P(X_n \leq x) \leq P(X \leq x) \quad \text{and} \quad \underline{\lim_{n\to\infty}}\, P(X_n \leq x) \geq P(X \leq x), \tag{6.1}$$

by developing inequalities. The bounds above together imply $\lim_{n\to\infty} P(X_n \le x) = P(X \le x)$ (Lemma C.1 in Appendix C). The proof of (6.1) uses the most fundamental of analytic techniques, namely giving ourselves an ε of room to maneuver and then taking $\varepsilon \searrow 0$ in the end.

First the upper bound on $\overline{\lim}$. Let $\varepsilon > 0$.

$$\begin{aligned} P(X_n \le x) &\le P(X_n \le x, |X_n - X| \le \varepsilon) + P(\,|X_n - X| > \varepsilon) \\ &\le P(X \le x + \varepsilon) + P(\,|X_n - X| > \varepsilon). \end{aligned}$$

Take $n \to \infty$ and use the assumption $X_n \xrightarrow{P} X$ to obtain

$$\begin{aligned} \varlimsup_{n\to\infty} P(X_n \le x) &\le P(X \le x + \varepsilon) + \lim_{n\to\infty} P(\,|X_n - X| > \varepsilon) \\ &= P(X \le x + \varepsilon). \end{aligned}$$

The inequality above is valid for all $\varepsilon > 0$. Hence we can let $\varepsilon \searrow 0$ while preserving the inequality, and appeal to the right-continuity of the cumulative distribution function (Theorem 2.20):

$$\varlimsup_{n\to\infty} P(X_n \le x) \le \lim_{\varepsilon\to 0+} P(X \le x + \varepsilon) = P(X \le x).$$

The first part of (6.1) has been established. To clarify, we wrote lim above in those situations where we already know that the limit exists.

The part done above used the right-continuity of F at x and did not appeal to the assumed continuity of F at x. The next lower bound proof will use the continuity. Start again with inequalities from additivity and monotonicity of probability:

$$\begin{aligned} P(X_n \le x) &\ge P(X_n \le x, |X_n - X| \le \varepsilon) \\ &\ge P(X \le x - \varepsilon, |X_n - X| \le \varepsilon) \\ &\ge P(X \le x - \varepsilon) - P(\,|X_n - X| > \varepsilon). \end{aligned}$$

As above, first $n \to \infty$:

$$\begin{aligned} \varliminf_{n\to\infty} P(X_n \le x) &\ge P(X \le x - \varepsilon) - \lim_{n\to\infty} P(\,|X_n - X| > \varepsilon) \\ &= P(X \le x - \varepsilon), \end{aligned}$$

and then $\varepsilon \searrow 0$:

$$\varliminf_{n\to\infty} P(X_n \le x) \ge \lim_{\varepsilon\to 0+} P(X \le x - \varepsilon) = \lim_{\varepsilon\to 0+} F(x - \varepsilon) = F(x).$$

We switched above from probability to F to highlight the use of the assumption of continuity of F at x. Both parts of (6.1) have now been verified and thereby the limit established. □

Distributional limits occupy a central role in probability as descriptions of the error in stronger limit theorems. In a typical situation Y_n is a sequence of random variables, c is a constant, and $Y_n \to c$ in probability or almost surely. The difference $Y_n - c$ shrinks as $n \to \infty$, but at the same time approximates a particular type of probability distribution. To observe this distribution we zoom closer into the difference by magnifying $Y_n - c$. This means studying the distributional limit $n^\alpha(Y_n - c) \xrightarrow{d} Z$ for some positive power α. The distribution of Z and the value of α depend on the situation. The limit can be turned around into an approximation:

for large n, $Y_n \overset{d}{\approx} c + n^{-\alpha} Z$ where $\overset{d}{\approx}$ means that the probability distributions of the left and right side are close. This approximation can be of practical use when Y_n is complicated for large n while Z is relatively simple.

In Section 6.2 Y_n is the maximum of independent random variables. In Section 6.4 the limit $Y_n \to c$ is the strong law of large numbers and the attendant limit in distribution is the *central limit theorem.*

6.2. Limit distribution of the maximum of i.i.d. random variables

This section looks at limit distributions associated with maxima of i.i.d. random variables. The starting point for studying a maximum is the following observation. If $X_1, X_2, \dots, X_n$ are i.i.d. then for any $x \in \mathbb{R}$

$$\begin{aligned} P\{\max(X_1, \dots, X_n) \le x\} &= P(X_1 \le x, \dots, X_n \le x) \\ &= \prod_{k=1}^{n} P(X_k \le x) = P(X_1 \le x)^n. \end{aligned} \tag{6.2}$$

Theorem 6.6. *Let* $X_1, X_2, X_3, \dots$ *be i.i.d.* $\mathrm{Unif}(0,1)$ *random variables. Let* $M_n = \max(X_1, \dots, X_n)$ *be the maximum of the first n random numbers. Then* $M_n \to 1$ *almost surely. Furthermore, we have the weak limit* $n(1 - M_n) \xrightarrow{d} Z$ *where* $Z \sim \mathrm{Exp}(1)$.

Proof. The almost sure limit $M_n \to 1$ follows from the Borel-Cantelli lemma (Exercise 5.6).

Let us observe how we are led to the exponential limit in distribution. By (6.2) for $x \in (0,1)$

$$P(M_n \le x) = P(X_1 \le x)^n = x^n.$$

For a fixed $x \in (0,1)$, $x^n \to 0$ and this just tells us that $M_n \xrightarrow{P} 1$. To get something nontrivial we take x closer and closer to 1 as $n \to \infty$. The exponential limit $(1 - \frac{y}{n})^n \to e^{-y}$ gives us a hint. Setting $x = 1 - \frac{y}{n}$ with $y \ge 0$ gives

$$\lim_{n\to\infty} P(M_n \le 1 - \tfrac{y}{n}) = \lim_{n\to\infty} (1 - \tfrac{y}{n})^n = e^{-y}.$$

By rearranging the terms we convert this into a limit of the cumulative distribution function of $n(1 - M_n)$. For $y \ge 0$,

$$\begin{aligned} \lim_{n\to\infty} P\{n(1 - M_n) \le y\} &= \lim_{n\to\infty} P(M_n \ge 1 - \tfrac{y}{n}) \\ &= \lim_{n\to\infty} \bigl(1 - P(M_n < 1 - \tfrac{y}{n})\bigr) \\ &= \lim_{n\to\infty} \bigl(1 - (1 - \tfrac{y}{n})^n\bigr) = 1 - e^{-y}. \end{aligned}$$

(Since we are dealing with absolutely continuous random variables, the probabilities $P(M_n < 1 - \frac{y}{n})$ and $P(M_n \le 1 - \frac{y}{n})$ are the same.)

For $y < 0$ the limit above is zero because $P(n(1 - M_n) < 0) = 0$ for all n. Altogether we have shown that the cumulative distribution function of $n(1 - M_n)$ converges to the cumulative distribution function of the Exp(1) distribution. □

The next example is different in that no additional scaling is required, only centering.

Example 6.7 (Distributional limit of the maximum of i.i.d. exponentials)**.** Let $X_1, X_2, \dots$ be i.i.d. Exp(1) random variables and set $M_n = \max(X_1, \dots, X_n)$. How fast does M_n grow?

We use a similar strategy as before. For $x_n > 0$ we have

$$P(M_n \le x_n) = P(X_1 \le x_n)^n = (1 - e^{-x_n})^n.$$

We would like to choose x_n so that $(1 - e^{-x_n})^n$ converges as $n \to \infty$ to a nonzero limit. We can do that by choosing x_n so that e^{-x_n} is equal to $\frac{y}{n}$ for some $y > 0$. In that case the limit is e^{-y}. Setting now $x_n = -\log(y/n)$ for $y > 0$ we get

$$\lim_{n\to\infty} P(M_n \le -\log(y/n)) = e^{-y}.$$

To get a limit in distribution, we would like to see the limit of cumulative distribution functions. Note that

$$P(M_n \le -\log(y/n)) = P(M_n - \log n \le -\log y)$$

Now introducing $z = -\log y$ we get $P(M_n - \log n \le -\log y) = P(M_n - \log n \le z)$ and

$$\lim_{n\to\infty} P(M_n - \log n \le z) = e^{-e^{-z}}. \tag{6.3}$$

Here z can be any real number. The function $F(z) = e^{-e^{-z}}$ is the cumulative distribution function of the *standard Gumbel distribution.* (Exercise 6.5 asks you to check that F is a distribution function.)

Thus the distribution of $M_n - \log n$ converges weakly to the standard Gumbel distribution. △

6.3. Gaussian distribution

In preparation for the central limit theorem, we introduce the Gaussian, or normal, distribution.

Definition 6.8. Let μ be real and $\sigma > 0$. A random variable X has the **normal distribution with mean** μ **and variance** σ^2 if X has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{6.4}$$

on the real line. Abbreviate this by $X \sim \mathcal{N}(\mu, \sigma^2)$.

We proceed to verify that f in (6.4) is a probability density with mean μ and and variance σ^2. We do this by studying the most important special case.

Definition 6.9. A random variable Z has the **standard normal distribution** (also called **standard Gaussian distribution**) if Z has density function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{6.5}$$

on the real line. Abbreviate this by $Z \sim \mathcal{N}(0, 1)$.

The next theorem checks that φ is a genuine probability density function.

Theorem 6.10.

$$\int_{-\infty}^{\infty} e^{-s^2/2}\, ds = \sqrt{2\pi}.$$

Proof. Compute the square of the integral as a double integral and switch to polar coordinates.

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-s^2/2}\, ds\right)^2 &= \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \cdot \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2-y^2/2}\, dx\, dy = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta \\ &= \int_0^{2\pi} \left(-e^{-r^2/2}\Big|_{r=0}^{r=\infty}\right) d\theta = \int_0^{2\pi} d\theta = 2\pi. \end{aligned}$$

□

Figure 1 shows the plot of the probability density of the standard normal distribution, the familiar bell shaped curve.

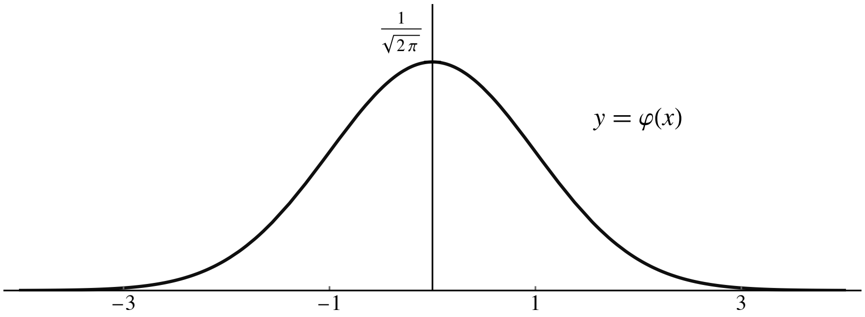

Figure 1. The probability density function φ of the standard normal distribution.

The standard normal distribution has its own notation: as above we write φ for the standard normal density function and

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-s^2/2}\, ds, \qquad x \in \mathbb{R}, \tag{6.6}$$

for the standard normal cumulative distribution function. Φ is continuous and strictly increasing, and satisfies $0 < \Phi(x) < 1$ for all $x \in \mathbb{R}$. See Figure 2.

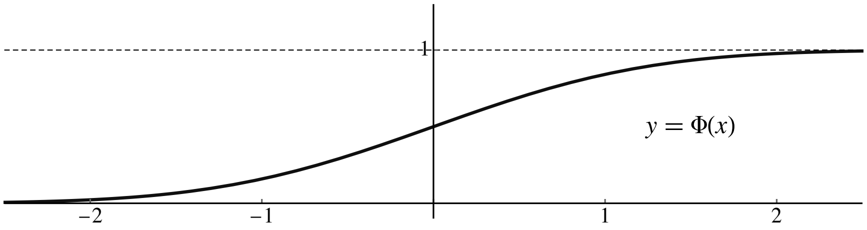

Figure 2. The cumulative distribution function Φ of the standard normal distribution.

There is no explicit antiderivative for φ. To find numerical values of $\Phi(x)$, we turn to the table in Appendix E. This table gives the values $\Phi(x)$ for $0 \le x \le 3.49$ accurate to four decimal digits. For larger x, $\Phi(x)$ will be closer than 0.0002 to 1. For negative values we use symmetry (see Figure 3). Since $\varphi(x) = \varphi(-x)$ we have

$$\Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-s^2/2}\, ds = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-s^2/2}\, ds = 1 - \Phi(x). \tag{6.7}$$

For example, to find $\Phi(-1.7)$, look up $\Phi(1.7) \approx 0.9554$ in the table and use $\Phi(-1.7) = 1 - \Phi(1.7) \approx 0.0446$.

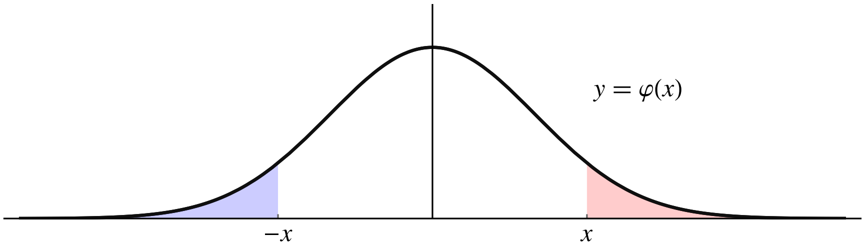

Figure 3. The symmetry of the function φ. The shaded blue area equals $\Phi(-x)$, the shaded red area equals $1 - \Phi(x)$, and the two are equal.

Example 6.11. Let $Z \sim \mathcal{N}(0, 1)$. Find the numerical value of $P(-1 \le Z \le 1.5)$.

$$\begin{aligned} P(-1 \le Z \le 1.5) &= \Phi(1.5) - \Phi(-1) = \Phi(1.5) - (1 - \Phi(1)) \\ &\approx 0.9332 - (1 - 0.8413) = 0.7745. \end{aligned}$$

The second to last step used the table in Appendix E. The answer is just an approximation of the probability because the values in the table are themselves decimal approximations of the exact values. △

Example 6.12. Find $z > 0$ so that a standard normal random variable Z has approximately $2/3$ probability of being in the interval $(-z, z)$.

$$P(-z < Z < z) = \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1.$$

Thus we need z for which $\Phi(z) = \frac{1}{2}(1 + \frac{2}{3}) = \frac{5}{6} \approx 0.833$. From the table in Appendix E we find $\Phi(0.96) = 0.8315$ and $\Phi(0.97) = 0.8340$. So $z = 0.97$ is a good approximation. We could use linear interpolation to get a slightly better approximation, but it is not important here. △

Theorem 6.13. *Let $Z \sim \mathcal{N}(0, 1)$. Then $E(Z) = 0$ and $\mathrm{Var}(Z) = E(Z^2) = 1$.*

Proof. By definition

$$E(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\, dx.$$

The first point is that this improper integral converges so the integral is well-defined. This is tested by integrating the absolute value of the integrand:

$$\int_{-\infty}^{\infty} |x| e^{-x^2/2}\, dx = 2 \int_{0}^{\infty} x e^{-x^2/2}\, dx = 2.$$

We used the symmetry of the integrated function. The last integral was evaluated above in the proof of Theorem 6.10. Then the value $E(Z) = 0$ follows from a general fact about integration: if f is an odd function, which means that $f(-x) = -f(x)$, then

$$\int_{-a}^{a} f(x)\,dx = 0$$

for any $a \in [0, \infty]$, as long as the integral is well-defined. Applying this to the odd function $f(x) = xe^{-x^2/2}$ gives $E(Z) = 0$.

For the mean square we integrate by parts.

$$\begin{aligned} E(Z^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x \cdot (-xe^{-x^2/2})\,dx \\ &= -\frac{1}{\sqrt{2\pi}} \left\{ xe^{-x^2/2} \Big|_{x=-\infty}^{x=\infty} - \int_{-\infty}^{\infty} e^{-x^2/2}\,dx \right\} \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1. \end{aligned}$$

In the second to last equality we used the limit $\lim_{x\to\pm\infty} xe^{-x^2/2} = 0$, while in the last equality we integrate the standard normal density function over the entire real line. □

Given real μ and $\sigma > 0$, let $Z \sim \mathcal{N}(0, 1)$ and define

$$X = \sigma Z + \mu.$$

Then $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The cumulative distribution function of X is

$$F(x) = P(X \le x) = P(\sigma Z + \mu \le x) = P(Z \le \tfrac{x-\mu}{\sigma}) = \Phi(\tfrac{x-\mu}{\sigma}). \tag{6.8}$$

The probability density function of X is obtained by differentiating F:

$$f(x) = F'(x) = \frac{d}{dx}\left[\Phi(\tfrac{x-\mu}{\sigma})\right] = \frac{1}{\sigma}\varphi(\tfrac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{6.9}$$

This is exactly the function in (6.4) which has now been shown to be a probability density function. Since X defined above has mean μ and variance σ^2 and has been identified as a $\mathcal{N}(\mu, \sigma^2)$ random variable, it follows that μ and σ^2 are the mean and variance of the $\mathcal{N}(\mu, \sigma^2)$ distribution.

The argument above generalizes to show that affine transformations $x \mapsto ax+b$ preserve the class of normal distributions, stated in the next theorem. Exercise 6.6 asks for the proof.

Theorem 6.14. *Let μ be real, $\sigma > 0$, and $X \sim \mathcal{N}(\mu, \sigma^2)$.*

(i) *Let $a \ne 0$, b real, and $Y = aX + b$. Then $Y \sim \mathcal{N}(a\mu + b,\, a^2\sigma^2)$.*

(ii) *$Z = \frac{X-\mu}{\sigma}$ is a standard normal random variable.*

The second part of the theorem above is used to evaluate probabilities of general normal random variables, by turning them into probabilities of standard normal random variables.

Example 6.15. Suppose $X \sim \mathcal{N}(-3, 4)$. Find the probability $P(X \le -1.7)$.

Since $Z = \frac{X-(-3)}{2} \sim \mathcal{N}(0,1)$ we have

$$\begin{aligned} P(X \le -1.7) &= P\left(\frac{X-(-3)}{2} \le \frac{-1.7-(-3)}{2} \right) = P(Z \le 0.65) \\ &= \Phi(0.65) \approx 0.7422, \end{aligned}$$

with the last equality from the table in Appendix E. △

Example 6.16. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. What is the probability that the observed value of X deviates from μ by more than 2σ? With $Z = \frac{X-\mu}{\sigma}$ (a standard normal) we have

$$\begin{aligned} P(\,|X-\mu| > 2\sigma) &= P(X < \mu - 2\sigma) + P(X > \mu + 2\sigma) \\ &= P\Big(\frac{X-\mu}{\sigma} < -2 \Big) + P\Big(\frac{X-\mu}{\sigma} > 2 \Big) \\ &= P(Z < -2) + P(Z > 2) = 2\big(1 - P(Z \le 2)\big) \\ &= 2\big(1 - \Phi(2)\big) \approx 2(1 - 0.9772) = 0.0456. \end{aligned}$$

This can be remembered as a rule of thumb: a normal random variable is within two standard deviations of its mean with probability over 95%. △

6.4. Central limit theorem

Theorem 6.17 (Central limit theorem)**.** *Suppose $X_1, X_2, X_3, \dots$ are i.i.d. random variables with finite mean $E[X_1] = \mu$ and finite variance $\mathrm{Var}(X_1) = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then for any fixed $-\infty \le a \le b \le \infty$ we have*

$$\lim_{n\to\infty} P\bigg(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b \bigg) = \Phi(b) - \Phi(a) = \int_a^b \tfrac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy. \tag{6.10}$$

Limit (6.10) remains valid if one or both of the inequalities inside the probabilities are switched from $\le$ to $<$.

We can relate statement (6.12) to the discovery in (5.18) of the critical scale $n^{-1/2}$ as follows. Limit (6.12) says that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \overset{d}{\approx} Z$$

where $Z \sim \mathcal{N}(0,1)$ and $\overset{d}{\approx}$ means now approximately equal in distribution. Rearranging gives

$$\frac{S_n}{n} \overset{d}{\approx} \mu + \frac{\sigma}{\sqrt{n}} Z.$$

This formulation decomposes S_n/n (approximately) into its mean μ plus a stochastic fluctuation that is normal and has order of magnitude $1/\sqrt{n}$.

In current textbooks the central limit theorem is typically proved with Fourier analysis (see for example [**Dur10**]). In Section 6.6 below we prove the case for Bernoulli variables with Stirling's approximation of the factorial.

Error bound in the central limit theorem.

An early version of the following theorem was proved by Berry in 1941 and Esseen in 1942.

Theorem 6.18. *Suppose that $X_1, X_2, X_3, \ldots$ are i.i.d. random variables with $E[X_1] = \mu$ and $\operatorname{Var}(X_1) = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then for any $x \in \mathbb{R}$ and positive integer n we have*

$$\left| P\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \le x \right) - \Phi(x) \right| \le \frac{3E[|X_1 - \mu|^3]}{\sigma^3 \sqrt{n}}. \tag{6.11}$$

6.5. Normal approximation of the binomial

In this section we use the central limit theorem to approximate binomial probabilities. We begin by stating the special of the central limit theorem. Fix $0 < p < 1$ and let $\{X_k\}_{k \ge 1}$ be i.i.d. Bernoulli variables with mean p. That is, the X_ks are independent and each has the probability mass function $P(X_k = 1) = p$ and $P(X_k = 0) = 1 - p$. Then $S_n = X_1 + \cdots + X_n$ is a $\operatorname{Bin}(n, p)$ random variable with mean $ES_n = np$ and variance $\operatorname{Var}(S_n) = n\sigma^2 = np(1-p)$.

Theorem 6.19 (Central limit theorem for Bernoulli random variables)**.** *Let $0 < p < 1$ be fixed and suppose that $S_n \sim \operatorname{Bin}(n, p)$. Then for any fixed $-\infty \le a \le b \le \infty$,*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b \right) = \Phi(b) - \Phi(a) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx. \tag{6.12}$$

Another name for Theorem 6.19 is the *de Moivre-Laplace theorem*, and it goes back to the early 1700s. We use this theorem as an imprecise approximation of binomial probabilities with Gaussian probabilities:

$$P\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b \right) \approx \Phi(b) - \Phi(a) \tag{6.13}$$

A commonly used rule of thumb is that the approximation is good if $np(1-p) > 10$.

Example 6.20. A fair coin is flipped 10,000 times. Estimate the probability that the number of heads is between 4850 and 5100.

Let S denote the number of heads observed. Then $S \sim \operatorname{Bin}(10{,}000, \frac{1}{2})$ with mean $E(S) = 10{,}000 \cdot \frac{1}{2} = 5000$ and variance $\operatorname{Var}(S) = 10{,}000 \cdot \frac{1}{2} \cdot \frac{1}{2} = 2500$. The probability in question is $P(4850 \le S \le 5100)$. Centering with the mean and dividing by the standard deviation transforms the probability as follows:

$$\begin{aligned} P(4850 \le S \le 5100) &= P\left(\frac{4850 - 5000}{\sqrt{2500}} \le \frac{S - 5000}{\sqrt{2500}} \le \frac{5100 - 5000}{\sqrt{2500}} \right) \\ &= P\left(-3 \le \frac{S - 5000}{50} \le 2 \right). \end{aligned}$$

So far everything is precise. Now we continue with approximation

$$\begin{aligned} P\left(-3 \le \frac{S - 5000}{50} \le 2 \right) &\approx \Phi(2) - \Phi(-3) \\ &\approx 0.9772 - (1 - 0.9987) = 0.9759. \end{aligned}$$

We actually made two approximations. First we approximated the binomial with the normal distribution, and then we approximated the cumulative distribution function of the standard normal using the table in the appendix.

What was the benefit of the approximation? The exact probability is

$$P(4850 \le S \le 5100) = \sum_{k=4850}^{5100} \binom{10,000}{k} 2^{-10,000}.$$

With modern computational tools it is easy to evaluate this expression. It is approximately 0.9765. Our approximate value 0.9759 is within 0.06% of the truth. The normal approximation is a quick way to get a number that is often good enough. △

Continuity correction.

A random variable $S_n \sim \mathrm{Bin}(n, p)$ takes only integer values. Thus, if k_1, k_2 are integers then

$$P(k_1 \le S_n \le k_2) = P(k_1 - 1/2 \le S_n \le k_2 + 1/2)$$

since the interval $[k_1, k_2]$ contains exactly the same integers as $[k_1 - 1/2, k_2 + 1/2]$. It turns out that if we apply the normal approximation to the modified interval $[k_1 - 1/2, k_2 + 1/2]$ we get a slightly better approximation of the exact binomial probability. Switching from $[k_1, k_2]$ to $[k_1 - 1/2, k_2 + 1/2]$ is called the *continuity correction.* It can be important if k_1, k_2 are close to each other or $np(1-p)$ is not very large.

Example 6.21. Roll a fair die 720 times. Estimate the probability that we have exactly 113 sixes.

Denote the number of sixes by S. This is a $\mathrm{Bin}(720, \frac{1}{6})$ distributed random variable with mean 120 and variance $720 \cdot \frac{1}{6} \cdot \frac{5}{6} = 100$. To estimate $P(S = 113)$ we write it as

$$P(S = 113) = P(112.5 \le S \le 113.5).$$

The normal approximation now gives

$$\begin{aligned} P(S = 113) &\approx \Phi\left(\frac{113.5 - 120}{10}\right) - \Phi\left(\frac{112.5 - 120}{10}\right) \\ &= \Phi(-0.65) - \Phi(-0.75) \approx (1 - 0.7422) - (1 - 0.7734) = 0.0312. \end{aligned}$$

The exact probability (calculated using computer) is

$$P(S = 113) = \binom{720}{113} \frac{5^{607}}{6^{720}} \approx 0.0318.$$

Our estimate is within 2% of the actual value which is an excellent approximation. Note that a mindless normal approximation without the continuity correction gives an absurd estimate, namely zero:

$$\begin{aligned} P(S = 113) &= P\left(\frac{113 - 120}{10} \le \frac{S - 120}{10} \le \frac{113 - 120}{10}\right) \\ &= P(-0.7 \le \tfrac{S-120}{10} \le -0.7) \approx \Phi(-0.7) - \Phi(-0.7) = 0. \end{aligned}$$

In this example $np(1-p) = 100$ so we are well within range of the rule of thumb of the normal approximation. △

Example 6.22 (Continuation of Example 6.20)**.** When the number of trials is large, the continuity correction does not make much of a difference. We illustrate this by applying the continuity correction to Example 6.20. A fair coin is flipped 10,000 times and we want the probability that we observe between 4850 and 5100 heads.

$$\begin{aligned}P(4850 \le S \le 5100) &= P(4849.5 \le S \le 5100.5)\\ &= P\left(\frac{4849.5-5000}{\sqrt{2500}} \le \frac{S-5000}{\sqrt{2500}} \le \frac{5100.5-5000}{\sqrt{2500}}\right)\\ &= P\left(-3.01 \le \frac{S-5000}{50} \le 2.01\right)\\ &\approx \Phi(2.01) - \Phi(-3.01)\\ &\approx 0.9778 - (1-0.9987) = 0.9765.\end{aligned}$$

The estimate improved slightly, but it was already very good before. △

Confidence intervals.

Suppose we have a biased coin and we do not know the true probability p that it lands on heads. How can we estimate p? The law of large numbers suggests a natural approach: flip the coin a large number n times, count the number S_n of heads, and take the observed frequency $\widehat{p} = \frac{S_n}{n}$ as the estimate for p. In practice we cannot flip the coin forever to get an accurate estimate. Can we estimate the error of the approximation for a finite n?

Let us see if normal approximation can say anything about this problem. We estimate the probability that the error $|\widehat{p} - p|$ is bounded by some small margin of error ε. First rearrange the inequality inside the probability and standardize:

$$\begin{aligned}P(|\widehat{p}-p| < \varepsilon) &= P\left(\left|\tfrac{S_n}{n} - p\right| < \varepsilon\right) = P\left(-n\varepsilon < S_n - np < n\varepsilon\right)\\ &= P\left(-\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \tfrac{S_n-np}{\sqrt{np(1-p)}} < \tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right).\end{aligned}$$

Up to this point we have used only algebra. Now comes the normal approximation:

$$\begin{aligned}P\left(-\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} < \tfrac{S_n-np}{\sqrt{np(1-p)}} < \tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\right) &\approx \Phi\Big(\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\Big) - \Phi\Big(-\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\Big)\\ &= 2\Phi\Big(\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\Big) - 1.\end{aligned}$$

We do not know the value of p (that is the whole point!) so how can we evaluate the last term? We can get a lower bound for it that works for all p. The maximum value of $p(1-p)$ is $1/4$ which is achieved at $p = 1/2$. (Check this with calculus.) Consequently $\frac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}} \ge 2\varepsilon\sqrt{n}$. Since the function Φ is increasing,

$$2\Phi\Big(\tfrac{\varepsilon\sqrt{n}}{\sqrt{p(1-p)}}\Big) - 1 \;\ge\; 2\Phi(2\varepsilon\sqrt{n}\,) - 1.$$

Combining the steps above gives the following lower bound on the probability, valid for all p, with the usual caveats about the validity of the normal approximation:

$$P(\,|\widehat{p}-p|<\varepsilon) \;\geq\; 2\Phi\big(2\varepsilon\sqrt{n}\,\big)-1. \tag{6.14}$$

Inequality (6.14) can be used to answer several types of questions. First, we might want to know how many trials are needed to reach a prescribed level of confidence in the estimate.

Example 6.23. How many times should we flip a coin with unknown success probability p so that the estimate $\widehat{p}=S_n/n$ is within 0.05 of the true p, with probability at least 0.99?

Using (6.14), to ensure a lower bound of at least 0.99 we need

$$P(\,|\widehat{p}-p|<\varepsilon)\geq 2\Phi\big(2\varepsilon\sqrt{n}\big)-1\geq 0.99.$$

This last inequality is satisfied if

$$\Phi\big(2\varepsilon\sqrt{n}\,\big)\geq 0.995.$$

From the table for Φ the last inequality is equivalent to

$$2\varepsilon\sqrt{n}\geq 2.58 \quad \text{which is the same as} \quad n\geq\frac{2.58^2}{4\varepsilon^2}=\frac{2.58^2}{4\cdot 0.05^2}\approx 665.64.$$

In the last step we used the value $\varepsilon=0.05$ that was given in the problem statement.

The conclusion is that if we flip the coin 666 times, the estimate $\widehat{p}$ is within 0.05 of the real p, with probability at least 0.99. △

Another task is to find the confidence interval around $\widehat{p}$ that captures the true p, with a given (high) probability. The $100r\%$ *confidence interval* for the unknown success probability p is given by $(\widehat{p}-\varepsilon,\widehat{p}+\varepsilon)$ where ε is chosen to satisfy $P(|\widehat{p}-p|<\varepsilon)\geq r$. In other words, the random interval $(\widehat{p}-\varepsilon,\widehat{p}+\varepsilon)$ contains the true p with probability at least r.

Example 6.24. We repeat a trial 1000 times and observe 450 successes. Find the 95% confidence interval for the true success probability p.

This time n is given and we look for ε such that $P(|\widehat{p}-p|<\varepsilon)\geq 0.95$. From (6.14) we need to solve the inequality $2\Phi\big(2\varepsilon\sqrt{n}\,\big)-1\geq 0.95$ for ε. First simplify and then turn to the Φ table:

$$\Phi\big(2\varepsilon\sqrt{n}\,\big)\geq 0.975 \quad\Longleftrightarrow\quad 2\varepsilon\sqrt{n}\geq 1.96 \quad\Longleftrightarrow\quad \varepsilon\geq\frac{1.96}{2\sqrt{1000}}\approx 0.031.$$

Thus if $n=1000$, then with probability at least 0.95 the random quantity $\widehat{p}$ satisfies $|\widehat{p}-p|<0.031$. If our observed ratio is $\widehat{p}=\frac{450}{1000}=0.45$, we say that the *95% confidence interval* for the true success probability p is $(0.45-0.031,0.45+0.031)=(0.419,0.481)$. △

Note carefully the terminology used in the example above. Once the experiment has been performed and 450 successes observed, $\widehat{p}=\frac{450}{1000}$ is no longer random. The true p is also not random since it is just a fixed parameter. Thus we can no longer say that "the true p lies in the interval $(\widehat{p}-0.031,\widehat{p}+0.031)=(0.419,0.481)$ with probability 0.95". That is why we say instead that $(\widehat{p}-0.031,\widehat{p}+0.031)=(0.419,0.481)$ is the *95% confidence interval* for the true p.

Remark 6.25 (Maximum likelihood estimator)**.** The use of $\widehat{p} = S_n/n$ as the estimate of the unknown success probability p was justified above by the law of large numbers. Here is an alternative justification. Once the outcome $S_n = k$ has been observed, we can use the value of the probability mass function of S_n to compare how likely the outcome k is under different parameter values p. We call it the *likelihood function* $L(p) = P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$, which is a function of p with fixed k and n.

The value $\widehat{p}$ that maximizes $L(p)$ is the *maximum likelihood estimator* of p. This is the value of p that gives the outcome k the highest probability. The reader can do the calculus to check that $L(p)$ is maximized uniquely by the value $\widehat{p} = k/n$. Thus the maximum likelihood estimator of p is the same $\widehat{p} = S_n/n$ for which we derived confidence intervals above. △

Polling.

Polling means estimating public opinion by interviewing a sample of people. This leads naturally to confidence intervals. We discuss below some mathematical issues that arise. Creating polls that are genuinely representative of the larger population is a very difficult practical problem which we do not address at all.

Example 6.26. Suppose that the fraction of a population who like broccoli is p. We wish to estimate p. In principle we could record the preferences of every individual, but this would be slow and expensive. Instead we take a random sample: we choose randomly n individuals, ask each of them whether they like broccoli or not, and estimate p with the ratio $\widehat{p}$ of those who said yes. We would like to quantify the accuracy of this estimate.

When we take a poll we are actually sampling *without replacement* because we do not ask the same individual twice. Recall that it is sampling *with replacement* that leads to independent trials and a binomially distributed number of successes. Consequently the number of people who said yes to broccoli is not exactly $\text{Bin}(n, p)$ but a hypergeometric-distributed random variable. So strictly speaking, the approach developed above for estimating p for independent trials is not valid now.

However, if the sample size n is small compared to the size of the population, even if we sampled with replacement the chances of asking the same person twice would be small. Consequently sampling with and without replacement are very close to each other. We discuss this point in more detail below. With this justification we can pretend that the sample of the poll was taken with replacement and thereby results in a $\text{Bin}(n, p)$ random variable. Then we can use the techniques developed above for the binomial.

Continuing with the example, suppose we interviewed 100 people and 20 of them liked broccoli. Thus our estimate is $\widehat{p} = \frac{20}{100} = 0.20$. Let us find the 90% confidence interval for the true p.

With $n = 100$ and a desired confidence level of 0.90, we seek ε such that

$$P(\,|\widehat{p} - p| < \varepsilon) \geq 0.90.$$

By inequality (6.14) this can be achieved by making sure ε satisfies

$$\begin{aligned} &2\Phi\bigl(2\varepsilon\sqrt{n}\,\bigr) - 1 \geq 0.90 \quad \Longleftrightarrow \quad \Phi\bigl(2\varepsilon\sqrt{n}\,\bigr) \geq 0.95 \\ \Longleftrightarrow \quad &2\varepsilon\sqrt{n} \geq 1.645 \quad \Longleftrightarrow \quad 20\varepsilon \geq 1.645 \quad \Longleftrightarrow \quad \varepsilon \geq 0.082. \end{aligned}$$

The value 1.645 in the calculation above was chosen because, according to the table, $\Phi(1.64) = 0.9495$ and $\Phi(1.65) = 0.9505$. Thus the 90% confidence interval for p is $(0.20 - 0.082, 0.20 + 0.082) = (0.118, 0.282)$.

Suppose the broccoli growers who commissioned the poll come back and tell you that the error 0.082 is too large. How many more people do you need to interview to reduce the margin of error down to 0.05, still achieving the same 90% confidence level? This time we take $\varepsilon = 0.05$ and solve for n:

$$\begin{aligned} &2\Phi\bigl(2\varepsilon\sqrt{n}\,\bigr) - 1 \geq 0.90 \quad \Longleftrightarrow \quad \Phi\bigl(2\varepsilon\sqrt{n}\,\bigr) \geq 0.95 \\ \Longleftrightarrow \quad &2\varepsilon\sqrt{n} \geq 1.645 \quad \Longleftrightarrow \quad 2\cdot 0.05\sqrt{n} \geq 1.645 \quad \Longleftrightarrow \quad n \geq 16.45^2 \approx 270.6. \end{aligned}$$

Thus to reach margin of error 0.05 with confidence level 90% requires 271 trials. In other words, after interviewing 100 people, another 171 interviews are needed. △

Remark 6.27 (Confidence levels in political polls)**.** During election seasons we are bombarded with news of the following kind: "The latest poll shows that 44% of voters favor candidate Honestman, with a margin of error of 3 percentage points". This report gives the confidence interval of the unknown fraction p that favor Honestman, namely $(0.44 - 0.03, 0.44 + 0.03) = (0.41, 0.47)$. The level of confidence used to produce the estimate is usually omitted from news reports. This is no doubt partly due to a desire to avoid confusing technicalities. It is also a fairly common convention to set the confidence level at 95%, so it does not need to be stated explicitly. △

Binomial limit of the hypergeometric.

Consider sampling without replacement from a set with two types of items, type A and type B. Let N_A be the number of type A items, N_B the number of type B items, and $N = N_A + N_B$ the total number of items. We sample n items *without replacement* (with $n \leq N$) and denote the number of type A items in the sample by X. Then X has the hypergeometric distribution defined below.

Definition 6.28. Let $0 \leq N_A \leq N$ and $1 \leq n \leq N$ be integers. A random variable X has the **hypergeometric distribution** with parameters (N, N_A, n) if X has probability mass function

$$P(X = k) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}}, \qquad \text{for } k = 0, 1, \ldots, n. \tag{6.15}$$

Abbreviate this by $X \sim \text{Hypergeom}(N, N_A, n)$. △

The possible values k of X satisfy $\max(0, n - N_B) \leq k \leq \min(n, N_A)$. This is because X cannot be larger than N_A, and the number $n - X$ of type B items sampled cannot be larger than $N_B = N - N_A$. Formula (6.15) gives the correct probability for all $k \in \{0, 1, \ldots, n\}$ when we use the convention $\binom{a}{k} = 0$ for integers $k > a \geq 0$.

When the sample size is small relative to the population size, it is intuitively clear that sampling with and without replacement should be close, because even with replacement the probability of drawing the same object twice is small. This is made precise in the next theorem.

Theorem 6.29. *Sample n items without replacement from a set of N items, N_A of which are of type A. Let X denote the number of type A items in the sample. Keeping n fixed, let N and N_A tend to infinity in such a way that $\frac{N_A}{N} \to p \in (0,1)$. Then*

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

In other words, X converges in distribution to a $\mathrm{Bin}(n,p)$ *random variable.*

Proof. With the notation $(a)_b = a\cdot(a-1)\cdots(a-b+1)$, the hypergeometric probability mass function of X is

$$\begin{aligned} P(X=k) &= \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}} = \frac{\frac{(N_A)_k}{k!}\cdot\frac{(N-N_A)_{n-k}}{(n-k)!}}{\frac{(N)_n}{n!}} \\ &= \binom{n}{k}\frac{(N_A)_k\,(N-N_A)_{n-k}}{(N)_n}. \qquad (6.16) \end{aligned}$$

With fixed n and k, take the limit of the last ratio above, as $N, N_A \to \infty$ and $N_A/N \to p$.

$$\begin{aligned} \frac{(N_A)_k\,(N-N_A)_{n-k}}{(N)_n} &= \frac{N_A(N_A-1)\cdots(N_A-k+1)}{N(N-1)\cdots(N-k+1)} \\ &\qquad \cdot \frac{(N-N_A)(N-N_A-1)\cdots(N-N_A-n+k+1)}{(N-k)(N-k-1)\cdots(N-n+1)} \\ &= \left(\frac{N_A}{N}\right)^k \cdot \prod_{i=1}^{k}\frac{(1-\frac{i-1}{N_A})}{(1-\frac{i-1}{N})} \cdot \left(1-\frac{N_A}{N}\right)^{n-k} \cdot \prod_{i=k+1}^{n}\frac{(1-\frac{i-k-1}{N-N_A})}{(1-\frac{i-1}{N})} \\ &\longrightarrow p^k\cdot 1\cdot(1-p)^{n-k}\cdot 1. \end{aligned}$$

The products of error terms converge to 1 above because under our assumptions N, N_A and $N-N_A$ all converge to ∞. We have shown that $P(X=k)$ converges to $\binom{n}{k}p^k(1-p)^{n-k}$. □

Theorem 6.29 is valid also for $p=0$ and $p=1$. We excluded those cases only to avoid worrying about additional details in the proof.

The hypergeometric-to-binomial limit explains why we can use the normal approximation for sampling without replacement for a reasonably large population (as in Example 6.26). In that case the hypergeometric distribution is close to the binomial distribution, which we can approximate with the normal.

Comparison of normal and Poisson approximation of the binomial.

Here is a restatement of the Poisson approximation theorem with error bound. This theorem is proved in Section 9.1.

Theorem 6.30 (Poisson approximation of the binomial with error term). *Let $S \sim \text{Bin}(n,p)$ and $Y \sim \text{Poisson}(np)$. Then*

$$\sum_{k=0}^{\infty} \big|P(S=k) - P(Y=k)\big| \le 2np^2. \tag{6.17}$$

In particular cases the reader may wonder which approximation to use for a binomial: the normal approximation or the Poisson approximation. When $np(1-p) > 10$ the normal approximation should be pretty safe as long as a and b are not too close together, while if $2np^2$ is small then the Poisson approximation will work well. The next two examples compare the two approximations. You can see that the normal and Poisson approximations are quite different and it is usually fairly evident which one to apply.

Example 6.31. Let $X \sim \text{Bin}(10, \frac{1}{10})$. Compare Poisson and normal approximations of the probability $P(X \le 1)$. The exact value is

$$P(X \le 1) = P(X=0) + P(X=1) = (\tfrac{9}{10})^{10} + 10 \cdot \tfrac{1}{10} \cdot (\tfrac{9}{10})^9 \approx 0.7361.$$

Now $E[X] = np = 10 \cdot \frac{1}{10} = 1$ so the correct Poisson approximation to use is $Y \sim \text{Poisson}(1)$. We have $2np^2 = 20 \cdot (\frac{1}{10})^2 = 0.2$ so the Poisson approximation will be within 0.2 of the true probability. In fact, it is even better than that:

$$P(Y \le 1) = P(Y=0) + P(Y=1) = e^{-1} + e^{-1} \approx 0.7358,$$

which is within 0.001 of the exact value.

Next we see how the normal approximation performs.

First without continuity correction. With $E[X] = 1$ and $\text{Var}(X) = 10 \cdot \frac{1}{10} \cdot \frac{9}{10} = \frac{9}{10}$ normal approximation gives

$$P(X \le 1) = P\bigg(\frac{X-1}{\sqrt{9/10}} \le \frac{1-1}{\sqrt{9/10}}\bigg) = P\bigg(\frac{X-1}{\sqrt{9/10}} \le 0\bigg) \approx \Phi(0) = 0.5.$$

Since the number of trials is small, we employ the continuity correction to enhance the accuracy of the normal approximation.

$$\begin{aligned} P(X \le 1) = P\big(X \le \tfrac{3}{2}\big) &= P\bigg(\frac{X-1}{\sqrt{9/10}} \le \frac{\frac{3}{2}-1}{\sqrt{9/10}}\bigg) \\ &= P\bigg(\frac{X-1}{\sqrt{9/10}} \le 0.53\bigg) \approx \Phi(0.53) = 0.7019. \end{aligned}$$

The approximation is not as good as the Poisson. Since $np(1-p) = 9/10$ is far below 10, the inferior performance of the normal approximation is not a surprise. △

The next example reverses the roles: the Poisson approximation performs poorly while the normal approximation does well.

Example 6.32. Estimate the probability that 40 flips of a fair coin give exactly 20 tails.

Let S be the number of tails among 40 coin flips. Then $S \sim \text{Bin}(40, \frac{1}{2})$ with mean $\mu = 20$ and standard deviation $\sigma = \sqrt{40 \cdot (1/2) \cdot (1/2)} = \sqrt{10}$. A calculation with a computer gives the exact probability as

$$P(S=20)=\binom{40}{20}2^{-40}\approx 0.1254.$$

The normal approximation with the continuity correction gives

$$\begin{aligned}P(S=20)=P\big(19.5\le S\le 20.5\big)&=P\Big(\frac{19.5-20}{\sqrt{10}}\le\frac{S-20}{\sqrt{10}}\le\frac{20.5-20}{\sqrt{10}}\Big)\\&=P\Big(-0.16\le\frac{S-20}{\sqrt{10}}\le 0.16\Big)\approx\Phi(0.16)-\Phi(-0.16)\\&=2\Phi(0.16)-1=0.1272.\end{aligned}$$

The fit is pretty good.

For the Poisson approximation we compare S to $Y\sim\text{Poisson}(20)$. A calculation using a computer gives

$$P(Y=20)=\frac{e^{-20}20^{20}}{20!}\approx 0.089.$$

The Poisson is badly off the mark. The reason is simply that we are not in the regime of the law of rare events, because successes with probability 1/2 are not rare! We have $2np^2=2\cdot 40\cdot\frac{1}{4}=20$ and so the rigorous error bound is larger than one. This bad error bound already suggests that Poisson is not the right approximation. △

6.6. Proof of the central limit theorem for Bernoulli variables

In this section we prove the CLT for i.i.d. Bernoulli variables. The proof is technical but requires nothing beyond calculus. We restate the CLT specialized to Bernoulli variables.

Fix $0<p<1$ and let $\{X_k\}_{k\ge1}$ be i.i.d. Bernoulli variables with mean p. That is, the X_ks are independent and each has the probability mass function $P(X_k=1)=p$ and $P(X_k=0)=1-p$. Then $S_n=X_1+\cdots+X_n$ is a $\text{Bin}(n,p)$ random variable with mean $ES_n=np$ and variance $\text{Var}(S_n)=n\sigma^2=np(1-p)$.

Theorem 6.33 (Central limit theorem for Bernoulli random variables)**.** *Let $0<p<1$ be fixed and suppose that $S_n\sim\text{Bin}(n,p)$. Then for any fixed $-\infty\le a\le b\le\infty$,*

$$\lim_{n\to\infty}P\Big(a\le\frac{S_n-np}{\sqrt{np(1-p)}}\le b\Big)=\Phi(b)-\Phi(a)=\int_a^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx. \qquad (6.18)$$

The remainder of this section proves Theorem 6.19. We begin with a sketch and then fill in the details.

Sketch of the proof of the CLT for Bernoulli variables.

For ease of notation set $q = 1 - p$. Begin by writing the probability on the left-hand side of (6.18) in terms of the binomial probability mass function:

$$\begin{aligned} P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) &= P\big(np + a\sqrt{npq} \le S_n \le np + b\sqrt{npq}\,\big) \\ &= \sum_{np+a\sqrt{npq}\le k\le np+b\sqrt{npq}} \frac{n!}{(n-k)!k!}\,p^k q^{n-k}, \end{aligned} \tag{6.19}$$

where the sum is over integers k between $np + a\sqrt{npq}$ and $np + b\sqrt{npq}$.

The rest of the proof comes in two steps. (i) Approximation of the factorial functions in the terms above with an expression that is easier to analyze. (ii) Interpretation of the resulting sum as a Riemann approximation of the integral on the right-hand side of (6.18).

Factorials are approximated by Stirling's formula:

$$n! \sim n^n e^{-n}\sqrt{2\pi n} \quad \text{as } n \to \infty. \tag{6.20}$$

The symbol $\sim$ in the statement has a precise technical meaning, namely

$$a_n \sim b_n \quad \text{means that} \quad \frac{a_n}{b_n} \to 1 \quad \text{as } n \to \infty. \tag{6.21}$$

Stirling's formula is proved in Theorem D.3 in Appendix D.

Proof of Theorem 6.33.

We proceed to prove the limit in (6.12) by carrying out the two steps outlined in the previous sketch: (i) Approximation of factorials with Stirling's formula. (ii) Interpretation of the resulting sum as a Riemann approximation of an integral of the Gaussian density function. The proof requires some fairly technical asymptotics that are developed in Appendix D.

Set $q = 1 - p$. Set

$$\gamma_n = \frac{n!}{n^n e^{-n}\sqrt{2\pi n}}.$$

Stirling's formula says that $\gamma_n \to 1$. For the summation bounds below, abbreviate $s_n = np + a\sqrt{npq}$ and $t_n = np + b\sqrt{npq}$.

$$\begin{aligned} &P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = P\big(np + a\sqrt{npq} \le S_n \le np + b\sqrt{npq}\,\big) \\ &\quad= \sum_{s_n\le k\le t_n} \frac{n!}{(n-k)!k!}\,p^k q^{n-k} \\ &\quad= \sum_{s_n\le k\le t_n} \frac{\gamma_n}{\gamma_{n-k}\gamma_k}\cdot \frac{n^n e^{-n}\sqrt{2\pi n}}{(n-k)^{n-k}\,e^{-n+k}\,\sqrt{2\pi(n-k)}\cdot k^k\,e^k\,\sqrt{2\pi k}}\,p^k q^{n-k} \\ &\quad= \frac{1}{\sqrt{2\pi}} \sum_{s_n\le k\le t_n} \frac{\gamma_n}{\gamma_{n-k}\gamma_k}\cdot \frac{n^n\,\sqrt{n}\,p^k q^{n-k}}{(n-k)^{n-k}\,k^k\,\sqrt{k(n-k)}}. \end{aligned} \tag{6.22}$$

A key point for the asymptotics is that the range of k in the sum above satisfies

$$|k - np| \le A\sqrt{n} \qquad \text{for a fixed constant } A. \tag{6.23}$$

This allows us to write below

$$\frac{k-np}{n} = O(n^{-1/2}) \qquad \text{uniformly for all } k \text{ in the sum.} \tag{6.24}$$

We argue below that the factor $\frac{\gamma_n}{\gamma_{n-k}\gamma_k}$ inside the sum on line (6.22) converges to 1 uniformly over the range of k in the sum. Hence that part is insignificant. Take one of the main factors in the sum. Rearrange it into a product of quantities with powers k, $n-k$ and $1/2$, and then exponentiate:

$$\begin{aligned}
&\frac{n^n \sqrt{n}\, p^k q^{n-k}}{(n-k)^{n-k}\, k^k\, \sqrt{k(n-k)}} \\
&= \Big(\frac{n-k}{nq}\Big)^{-n+k} \Big(\frac{k}{np}\Big)^{-k} \Big(\frac{n-k}{nq}\Big)^{-1/2} \Big(\frac{k}{np}\Big)^{-1/2} \frac{1}{\sqrt{npq}} \\
&= \exp\Big\{ -(n-k)\log\frac{n-k}{nq} - k\log\frac{k}{np} - \tfrac12\log\frac{n-k}{nq} - \tfrac12\log\frac{k}{np} \Big\} \frac{1}{\sqrt{npq}}\,. \qquad (6.25)
\end{aligned}$$

To take advantage of the asymptotics (D.1) of the logarithm, rewrite the terms above as follows.

$$\begin{aligned}
(n-k)\log\frac{n-k}{nq} &= (n-k)\log\Big(1-\frac{k-np}{nq}\Big) \\
&= (n-k)\Big\{-\frac{k-np}{nq} - \frac{(k-np)^2}{2n^2q^2} + O(n^{-3/2})\Big\} \qquad (6.26)\\
&= -\frac{(n-k)(k-np)}{nq} - \frac{(n-k)(k-np)^2}{2n^2q^2} + O(n^{-1/2}).
\end{aligned}$$

Multiplying $O(n^{-3/2})$ above with k produces a term of order $O(n^{-1/2})$ because k is of order n.

Similarly for the second term in the exponential on line (6.25):

$$\begin{aligned}
k\log\frac{k}{np} &= k\log\Big(1+\frac{k-np}{np}\Big) = k\Big\{\frac{k-np}{np} - \frac{(k-np)^2}{2n^2qp^2} + O(n^{-3/2})\Big\} \\
&= \frac{k(k-np)}{np} - \frac{k(k-np)^2}{2n^2p^2} + O(n^{-1/2}).
\end{aligned} \tag{6.27}$$

To the last two terms in the exponential on line (6.25) apply (D.3):

$$\tfrac12\log\frac{n-k}{nq} = \tfrac12\log\Big(1-\frac{k-np}{nq}\Big) = O(n^{-1/2}) \tag{6.28}$$

and

$$\tfrac12\log\frac{k}{np} = \tfrac12\log\Big(1+\frac{k-np}{np}\Big) = O(n^{-1/2}). \tag{6.29}$$

Substitute (6.26)–(6.29) inside the braces on line (6.25), combine all the $O(n^{-1/2})$ error terms, and take $\frac{k-np}{n}$ and $\frac{(k-np)^2}{n}$ as common factors, to develop the quantity in braces as follows:

$$
\begin{aligned}
&-(n-k)\log\frac{n-k}{nq} - k\log\frac{k}{np} - \tfrac{1}{2}\log\frac{n-k}{nq} - \tfrac{1}{2}\log\frac{k}{np} \\
&\quad = \frac{(n-k)(k-np)}{nq} - \frac{k(k-np)}{np} + \frac{(n-k)(k-np)^2}{2n^2q^2} + \frac{k(k-np)^2}{2n^2p^2} + O(n^{-1/2}) \\
&\quad = \left(\frac{n-k}{q} - \frac{k}{p}\right)\frac{k-np}{n} + \left(\frac{n-k}{nq^2} + \frac{k}{np^2}\right)\frac{(k-np)^2}{2n} + O(n^{-1/2}) \\
&\quad = -\frac{(k-np)^2}{npq} + \left(1 + \frac{(q-p)(k-np)}{npq}\right)\frac{(k-np)^2}{2npq} + O(n^{-1/2}) \\
&\quad = -\frac{(k-np)^2}{npq} + \frac{(k-np)^2}{2npq} + \frac{(q-p)(k-np)}{npq}\cdot\frac{(k-np)^2}{2npq} + O(n^{-1/2}) \\
&\quad = -\frac{(k-np)^2}{2npq} + O(n^{-1/2}).
\end{aligned}
\tag{6.30}
$$

The last equality came from noting that $\frac{k-np}{n} = O(n^{-1/2})$ and $\frac{(k-np)^2}{n} = O(1)$.

We work our way back up. The expression on line (6.30) goes back inside the braces on line (6.25). Entire line (6.25) replaces the long fraction at the end of line (6.22). Thus we can record the approximation

$$
\begin{aligned}
&P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) \\
&\qquad = \frac{1}{\sqrt{2\pi npq}} \sum_{s_n \le k \le t_n} \frac{\gamma_n}{\gamma_{n-k}\gamma_k} \cdot \exp\left\{-\frac{(k-np)^2}{2npq} + O(n^{-1/2})\right\}
\end{aligned}
\tag{6.31}
$$

For the last step we switch to precise inequalities. The Stirling's formula corrections represented by the γ_k factors satisfy the uniform estimates of the next lemma.

Lemma 6.34. *There exists positive constants ε_n such that $\varepsilon_n \to 0$ as $n \to \infty$ and for all n we have the bounds*

$$
1 - \varepsilon_n \le \min_{s_n \le k \le t_n} \frac{\gamma_n}{\gamma_{n-k}\gamma_k} \le \max_{s_n \le k \le t_n} \frac{\gamma_n}{\gamma_{n-k}\gamma_k} \le 1 + \varepsilon_n. \tag{6.32}
$$

Proof. To be filled in. □

Approximation (6.31) gives the following upper and lower bounds for the probability. We write C for the positive constant associated with the $O(n^{-1/2})$ in the exponential, that is also uniform over the range of k in the sum.

$$\begin{aligned}&\frac{(1-\varepsilon_n)e^{-Cn^{-1/2}}}{\sqrt{2\pi}} \sum_{s_n\le k\le t_n} \exp\Big\{-\frac{(k-np)^2}{2npq}\Big\}\frac{1}{\sqrt{npq}}\\ &\qquad\le P\Big(a\le \frac{S_n-np}{\sqrt{npq}}\le b\Big) \qquad (6.33)\\ &\qquad\le \frac{(1+\varepsilon_n)e^{Cn^{-1/2}}}{\sqrt{2\pi}} \sum_{s_n\le k\le t_n} \exp\Big\{-\frac{(k-np)^2}{2npq}\Big\}\frac{1}{\sqrt{npq}}\end{aligned}$$

The sum that is in both the upper and lower bound is a Riemann sum over the interval $[a,b]$ for the normal density function. To make this explicit, define partition points

$$x_k = \frac{k-np}{\sqrt{npq}}.$$

The summation limits $np + a\sqrt{npq} \le k \le np + b\sqrt{npq}$ then specify exactly $a \le x_k \le b$, and the length of the subintervals is $\Delta x = \frac{1}{\sqrt{npq}}$. The upper and lower bounds from (6.33) above become

$$(1\pm\varepsilon_n)e^{\pm Cn^{-1/2}} \sum_{k:\, a\le x_k\le b} \frac{1}{\sqrt{2\pi}}\, e^{-x_k^2/2}\,\Delta x \;\underset{n\to\infty}{\longrightarrow}\; \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx.$$

The last limit is precisely the Riemann sum approximation of the integral. To summarize, we have shown that, for $-\infty < a < b < \infty$,

$$\lim_{n\to\infty} P\Big(a\le \frac{S_n-np}{\sqrt{npq}}\le b\Big) = \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx. \qquad (6.34)$$

The estimates along the way depended crucially on a and b being finite because we used repeatedly estimate (6.24). We need to do a little more work to verify that

$$\lim_{n\to\infty} P\Big(\frac{S_n-np}{\sqrt{npq}}\le b\Big) = \Phi(b) = \int_{-\infty}^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\,dx, \qquad (6.35)$$

in other words, that we can take $a = -\infty$. We shall prove this limit by showing that the inequality

$$\varlimsup_{n\to\infty} \Big| P\Big(\frac{S_n-np}{\sqrt{npq}}\le b\Big) - \Phi(b)\Big| \le \varepsilon \qquad (6.36)$$

holds for every $\varepsilon > 0$. Then this last limit above must actually be zero, which is exactly the statement we want.

To prove (6.36) for a given b, pick $c > 0$ large enough so that $-c < b$ and

$$\Phi(-c) + \big(1-\Phi(c)\big) < \varepsilon/2.$$

This possible since as a distribution function, Φ satisfies $\lim_{x\to\infty}\Phi(x) = 1$ and $\lim_{x\to-\infty}\Phi(x) = 0$. Estimate the left tail probability:

$$\begin{aligned}
P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big) &\le P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big) + P\Big(\frac{S_n - np}{\sqrt{npq}} > c\Big)\\
&= 1 - P\Big(-c \le \frac{S_n - np}{\sqrt{npq}} \le c\Big) \quad \underset{n\to\infty}{\to} \quad 1 - \big(\Phi(c) - \Phi(-c)\big) < \varepsilon/2.
\end{aligned}$$

The last limit above comes from the already proved case (6.34). This calculation tells us that

$$\overline{\lim_{n\to\infty}}\, P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big) < \varepsilon/2.$$

(Since we have not proved the existence of the limit above, we must use limsup.) Next we argue with the triangle inequality.

$$\begin{aligned}
&\bigg| P\Big(\frac{S_n - np}{\sqrt{npq}} \le b\Big) \;-\; \Phi(b)\bigg|\\
&= \bigg| P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big) + P\Big(-c \le \frac{S_n - np}{\sqrt{npq}} \le b\Big) \;-\; \big(\Phi(b) - \Phi(-c)\big) - \Phi(-c)\bigg|\\
&\le P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big) + \bigg| P\Big(-c \le \frac{S_n - np}{\sqrt{npq}} \le b\Big) \;-\; \big(\Phi(b) - \Phi(-c)\big)\bigg| + \Phi(-c).
\end{aligned}$$

Let $n \to \infty$ and use $\Phi(-c) < \varepsilon/2$. .

$$\begin{aligned}
&\overline{\lim_{n\to\infty}}\bigg| P\Big(\frac{S_n - np}{\sqrt{npq}} \le b\Big) \;-\; \Phi(b)\bigg|\\
&\quad\le \overline{\lim_{n\to\infty}}\, P\Big(\frac{S_n - np}{\sqrt{npq}} < -c\Big)\\
&\qquad\quad + \overline{\lim_{n\to\infty}}\bigg| P\Big(-c \le \frac{S_n - np}{\sqrt{npq}} \le b\Big) \;-\; \big(\Phi(b) - \Phi(-c)\big)\bigg| + \frac{\varepsilon}{2}\\
&\le \frac{\varepsilon}{2} + 0 + \frac{\varepsilon}{2} = \varepsilon.
\end{aligned}$$

We have verified (6.36) and thereby completed the proof of the central limit theorem for Bernoulli variables.

6.7. Further mathematical issues ♣

Weak convergence of probability distributions on the real line can be defined by a metric on probability distributions. The *Lévy metric* is defined for two cumulative distribution functions F and G by

$$\rho(F, G) = \inf\{\varepsilon > 0 : F(x - \varepsilon) - \varepsilon \le G(x) \le F(x + \varepsilon) + \varepsilon \;\forall x \in \mathbb{R}\}. \tag{6.37}$$

Then $F_n \overset{d}{\to} F$ if and only if $\rho(F_n, F) \to 0$. Under the metric ρ, the space of probability distributions on $\mathbb{R}$ is a complete, separable metric space. That is, Cauchy sequences converge and there is a countable dense set.

Exercises

Exercise 6.1. Let us write $X_n \xrightarrow{P} c$ if X_n converges in probability to a random variable X that is constant c with probability 1, that is, $P(X=c)=1$. Similarly, write $X_n \xrightarrow{d} c$ if X_n converges to a constant random variable c in distribution. Show that $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{P} c$. (Together with Theorem 6.5 this implies that if the limit is a constant random variable, then convergence in probability in equivalent to convergence in distribution.)

Exercise 6.2. For each positive integer n, let X_n be a *uniform* random variable on the set $\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\}$, by which we mean that $P(X_n = \frac{k}{n}) = \frac{1}{n}$ for each $k \in \{1, 2, \dots, n\}$. Let F_n be the cumulative distribution function of X_n. Find the cumulative distribution function F such that $F_n \xrightarrow{d} F$.

Exercise 6.3. Let $\{X_n\}_{n\ge 1}$ and X be $\mathbb{Z}_{\ge 0}$-valued random variables. Show that $X_n \xrightarrow{d} X$ is equivalent to the statement that $P(X_n = k) \to P(X = k)$ for all $k \in \mathbb{Z}_{\ge 0}$.

Exercise 6.4. Let X_n have cumulative distribution function F_n and X have cumulative distribution function F. Assume that $X_n \xrightarrow{d} X$ and that F is a continuous function. Show that then F_n converges to F uniformly on $\mathbb{R}$.

Recall that a sequence of functions g_n converges to g *uniformly on a set* S if $\lim_{n\to\infty} \sup_{x\in S} |g_n(x) - g(x)| = 0$.

Exercise 6.5. Check that the function $F(x) = e^{-e^{-x}}$ on $\mathbb{R}$ is a cumulative distribution function.

Exercise 6.6. Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$ with $a \ne 0$. Show that $Y \sim \mathcal{N}(a\mu + b,\, a^2\sigma^2)$.

Exercise 6.7. Let $\mu \in \mathbb{R}$, σ_n a sequence in $(0,\infty)$, and $X_n \sim \mathcal{N}(\mu, \sigma_n^2)$. Suppose σ_n converges to some $\sigma \in [0,\infty]$. For which $\sigma \in [0,\infty]$ is there a limit in distribution $X_n \xrightarrow{d} X$? Describe the limit.

Exercise 6.8. Let $\mu \in \mathbb{R}$, σ_n a sequence in $(0,\infty)$ such that $\sigma_n \to 0$, and $X_n \sim \mathcal{N}(\mu, \sigma_n^2)$. Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded continuous function. Show that $E[h(X_n)] \to h(\mu)$ as $n \to \infty$.

Hint. Let f_n be the density function of X_n. Split the integral

$$\big|E[h(X_n)] - h(\mu)\big| \le E|h(X_n) - h(\mu)| = \int_{-\infty}^{\infty} \big|h(x) - h(\mu)\big|\, f_n(x)\, dx$$

into three pieces $\int_{-\infty}^{\mu-\delta} + \int_{\mu-\delta}^{\mu+\delta} + \int_{\mu+\delta}^{\infty}$ for a small $\delta > 0$ and estimate the pieces separately.

Significance. Let $\varphi_n(x) = \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\{-\frac{x^2}{2\sigma_n^2}\}$. Then the exercise shows that

$$(h * \varphi_n)(\mu) = \int_{-\infty}^{\infty} h(x)\, \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = E[h(X_n)] \to h(\mu).$$

Thus the convolution $h * \varphi_n$ approximates h for any bounded continuous h, and so φ_n can be viewed as approximating the identity operator on functions. The sequence $\{\varphi_n\}$ is an example of an *approximate identity.*

Exercise 6.9. You flip a fair coin 10,000 times. Use the CLT to approximate the probability that the difference between the number of heads and number of tails is at most 100.

Exercise 6.10. Let D be the planar region bounded by the curve $y = x^2$, the vertical line $x = 1$, and the x-axis. Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \dots$ be an infinite sequence of independent random points, and each point (X_i, Y_i) is uniformly distributed on the region D.

(1) Explain how the random sequence $n^{-1} \sum_{i=1}^{n} X_i^2$ behaves as $n \to \infty$.

(2) Find sequences of real numbers a_n and b_n such that the random quantities

$$\frac{\sum_{i=1}^{n} X_i^2 - a_n}{b_n}$$

converge in distribution as $n \to \infty$, and identify the limit distribution by name. (There are two answers, but only one of them is interesting.)

Exercise 6.11. A pollster would like to estimate the fraction p of people in a population who intend to vote for a particular candidate. How large must a random sample be in order to be at least 95% certain that the fraction $\widehat{p}$ of positive answers in the sample is within 0.02 of the true p?

Exercise 6.12. A political interest group wants to determine what fraction $p \in (0, 1)$ of the population intends to vote for candidate A in the next election. 1,000 randomly chosen individuals are polled. 457 of these indicate that they intend to vote for candidate A. Find the 95% confidence interval for the true fraction p.

Exercise 6.13. Suppose 10 percent of households earn over 80,000 dollars a year, and 0.25 percent of households earn over 450,000. A random sample of 400 households has been chosen. In this sample, let X be the number of households that earn over 80,000, and let Y be the number of households that earn over 450,000. Use either the normal or the Poisson approximation, whichever is appropriate in either case, to find the simplest estimates you can for the probabilities $P(X \geq 48)$ and $P(Y \geq 2)$.

Exercise 6.14. Let $S \sim \text{Bin}(10^{10}, 10^{-6})$. Use the error bounds in (6.11) and (6.17) to observe that the normal and Poisson approximations of $P(S \leq 10^4 + 10^2)$ are roughly equal in accuracy. Explain what is going on.

Hint. Recall the property of Poisson distributions in Example 3.44.

Chapter 7

Generating functions

We introduce three generating functions for random variables and then work with one of them, namely the moment generating function.

7.1. Three generating functions

We have seen various ways to describe the distribution of a random variable:

- The full distribution: all probabilities of the form $P(X \in B)$
- the cumulative distribution function
- For discrete random variables: the probability mass function
- For absolutely continuous random variables: the probability density function

Generating functions provide an alternative way to describe probability distributions. They encode the distribution of a random variable into a function defined by an expectation.

Here are the definitions:

- The *probability generating function* of a random variable X is defined as the function $G_X(s) = Es^X$ (for whichever s this is defined).
- The *moment generating function* of a random variable X is defined as the function $M_X(t) = Ee^{tX}$ (for whichever t this is defined).
- The *characteristic function* of a random variable X is defined as the function $\phi_X(t) = Ee^{itX}$.
  Note: e^{itX} is a complex valued random variable. The expectation of a complex random variable Z is defined as $E\Re Z + iE\Im Z$. (You just take the expectation of real and imaginary parts separately.)

These functions are closely related: $G(e^t) = M(t)$ and ϕ is just an extension of M to the imaginary axis. However, in different situations some versions can be more useful than the others.

All of these objects have multivariate versions. If we need joint generating functions of $X_1, \dots, X_n$ then we need n variables, e.g. the probability generating function is $G(s_1, \dots, s_n) = E\left[\prod_{i=1}^n s_i^{X_i}\right]$.

We will now look at the various versions of generating functions separately.

Probability generating function. This is especially useful if the random variable X is nonnegative and integer valued. This is because

$$G_X(s) = Es^X = \sum_{k=0}^{\infty} P(X = k)s^k,$$

and this is a Taylor series where the coefficients are exactly the values of the probability mass function.

Simple properties:

- $G_X(s)$ is finite for $|s| \le 1$. This is because the coefficients of the Taylor series are between 0 and 1, so the radius of convergence is at least 1.
- $G_X(0) = P(X = 0)$, $G_X(1) = 1$.
- If we know $G_X(s)$ then we can find the PMF by looking at the value and the derivatives at 0.
$$\frac{d^k}{s^k} G_X(s)\Big|_{s=0} = k!P(X = k)$$
Thus $G_X(s)$ identifies the distribution if X is nonnegative and integer valued.
- Since $G'(s) = \sum_{k=1}^{\infty} k s^{k-1} P(X = k)$, we have $EX = G'(1)$ (if $EX < \infty$). Similar formulas work for higher order derivatives:
$$EX(X-1)\cdots(X-(k-1)) = G^{(k)}(1)$$
(with both side equal to ∞ if the expectation is infinite.)

Examples.

- Bernoulli(p):
$$G(s) = ps + 1 - p, \quad M(t) = pe^t + (1-p), \quad \phi(t) = pe^{it} + 1 - p$$
- Geometric(p)
$$G(s) = \sum_{k=1}^{\infty} s^k p q^{k-1} = \frac{ps}{1-qs}, \quad M(t) = \frac{pe^t}{1-qe^t}$$
- Poisson(λ)
$$G(s) = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda s - \lambda}, \quad M(t) = e^{\lambda(e^t - 1)}$$

Generating functions are widely used in other areas of mathematics as well. The generating function of a sequence $a_0, a_1, \dots$ is defined as the function $f(s) = \sum_{i=0}^{\infty} a_i s^i$. The probability generating function is then just the generating function of the sequence $a_k = P(X = k)$. (Note however that in the general case nothing guarantees that the generating function is well defined for any $s \neq 0$.)

Moment generating function. Simple properties:

- $M_X(0) = 1$, but it might happen that none of the other values are defined. If X is nonnegative then $M_X(t) < \infty$ for $t < 0$.
- If $M_X(t)$ is defined in a small neighborhood of 0 then the derivatives at zero are equal to the moments (if they exist).

  $$\frac{d}{dt} E e^{tX} = E(e^{tX} X), \qquad M'(0) = EX.$$

  (One would need to justify the fact that we can differentiate inside the expectation. If $M_X(t)$ is finite in a neighborhood of 0 then this is justified.) Same works for higher order derivatives (if the moments exist):

  $$M_X^{(n)}(0) = EX^n.$$

  (It's easier to get moments from the moment generating function.)
- If $M_X(t)$ is finite in $(-\varepsilon, \varepsilon)$ then the values of the function identify the distribution. In particular, if X and Y have moment generating functions that agree on $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$ then X and Y have the same distributions. (There is an inversion formula.)
- The moment generating function of $aX + b$ can be computed in terms of the MGF of X:

  $$M_{aX+b}(t) = E e^{(aX+b)t} = e^{bt} E e^{atX} = e^{bt} M_X(t).$$

Characteristic function. Simple properties:

- $\phi_X(0) = 1$, and $\phi_X(t)$ is defined for all t!! (Because e^{iXt} is bounded.)
- The derivatives at zero will produce the moments (if they exist).

  $$\frac{d}{dt} E e^{itX} = E(e^{itX} X), \qquad \phi'(0) = iEX.$$

  (One would again need to justify the fact that we can differentiate inside the expectation.) Same works for higher order derivatives (if the moments exist):

  $$\phi_X^{(n)}(0) = i^n EX^n.$$
- The characteristic function identifies the distribution. There is an inversion formula.

7.2. Moment generating function

We take a closer look at applications of the moment generating function to calculate moments and to identify distributions of sums. In the last part of this section we sketch a proof of the central limit theorem with moment generating functions. We begin with examples of moment generating functions of familiar named distributions.

Example 7.1 (Moment generating function of the Poisson distribution)**.** Let $X \sim \text{Poisson}(\lambda)$.

$$E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\,\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(e^t\lambda)^k}{k!} = e^{-\lambda}\cdot e^{e^t\lambda} = e^{\lambda(e^t-1)},$$

where we used the series expansion of the exponential function (see (C.6)). Thus for $X \sim \text{Poisson}(\lambda)$, we have $M_X(t) = e^{\lambda(e^t-1)}$ for all real t. △

Example 7.2 (Moment generating function of the normal distribution)**.** Let $Z \sim \mathcal{N}(0,1)$. To evaluate the integral we complete the square in the exponential.

$$\begin{aligned} E(e^{tZ}) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-x^2/2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2+tx-\frac{1}{2}t^2+\frac{1}{2}t^2}\,dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}\,dx = e^{t^2/2}. \end{aligned}$$

Notice how we immediately know the value of the last integral because we recognize the integrand as the density function of the $\mathcal{N}(t,1)$ distribution. We deduced that $M_Z(t) = e^{t^2/2}$ for $Z \sim \mathcal{N}(0,1)$.

To get the m.g.f. of a general normal random variable $X \sim \mathcal{N}(\mu,\sigma^2)$, write $X = \sigma Z + \mu$ and deduce

$$E(e^{tX}) = E(e^{t(\sigma Z+\mu)}) = e^{t\mu}E(e^{t\sigma Z}) = e^{t\mu}\cdot e^{t^2\sigma^2/2} = e^{\mu t+\sigma^2t^2/2}.$$

△

Example 7.3 (Moment generating function of the exponential distribution)**.** Let $X \sim \text{Exp}(\lambda)$. Then

$$E(e^{tX}) = \int_0^{\infty} e^{tx}\cdot\lambda e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}\,dx = \lambda\lim_{b\to\infty}\int_0^b e^{(t-\lambda)x}\,dx.$$

This is an improper integral whose value depends on t, so let us do it carefully. If $t = \lambda$,

$$E(e^{tX}) = \lambda\lim_{b\to\infty}\int_0^b 1\,dx = \lambda\lim_{b\to\infty} b = \infty.$$

Suppose $t \neq \lambda$.

$$E(e^{tX}) = \lambda\lim_{b\to\infty}\left(\frac{e^{(t-\lambda)x}}{t-\lambda}\Bigg|_{x=0}^{x=b}\right) = \lambda\lim_{b\to\infty}\frac{e^{(t-\lambda)b}-1}{t-\lambda} = \begin{cases}\infty & \text{if } t > \lambda,\\ \dfrac{\lambda}{\lambda-t} & \text{if } t < \lambda.\end{cases}$$

To summarize, the m.g.f. of the $\text{Exp}(\lambda)$ distribution is

$$M(t) = \begin{cases} \infty & \text{if } t \geq \lambda, \\ \dfrac{\lambda}{\lambda - t} & \text{if } t < \lambda. \end{cases}$$

△

The exponential example above shows that $M(t)$ can be infinite for some portion of t-values. In fact, it can happen that all values except $M(0) = 1$ are infinite.

Calculation of moments with the moment generating function.

Let $M(t) = E[e^{tX}]$ and consider the following calculation:

$$M'(t) = \frac{d}{dt} E[e^{tX}] = E\left[\frac{d}{dt} e^{tX}\right] = E[Xe^{tX}].$$

Substituting $t = 0$ gives the formula $M'(0) = E[X]$. The fact that we can move the differentiation inside the expectation is not self-evident, but we will not discuss the justification here. However, in the case that X takes only finitely many values this step is straightforward because the derivative of a sum is the sum of the derivatives. We have

$$\begin{aligned} M'(t) &= \frac{d}{dt} E[e^{tX}] = \frac{d}{dt} \sum_k e^{kt} P\{X = k\} = \sum_k \frac{d}{dt} e^{kt} P\{X = k\} \\ &= \sum_k k e^{kt} P\{X = k\} = E[Xe^{tX}]. \end{aligned} \tag{7.1}$$

Setting $t = 0$ gives the identity

$$M'(0) = \sum_k k P\{X = k\} = E[X].$$

Returning to the general case, we can continue to differentiate as many times as we please by taking the derivative inside the expectation. Write $M^{(n)}$ for the nth derivative of the function M.

$$M^{(n)}(t) = \frac{d^n}{dt^n} E[e^{tX}] = E\left[\frac{d^n}{dt^n} e^{tX}\right] = E[X^n e^{tX}].$$

Taking $t = 0$ gives the following formula.

Theorem 7.4. *When the moment generating function $M(t)$ of a random variable X is finite in an interval around the origin, then all moments of X are finite and are given by*

$$E(X^n) = M^{(n)}(0).$$

Example 7.5 (Moments of the Bernoulli distribution)**.** Let X be Bernoulli with parameter p. Then $M_X(t) = (1 - p) + pe^t$. Therefore, $M_X^{(n)}(t) = pe^t$ for all $n \geq 1$. Thus, $E[X^n] = M_X^{(n)}(0) = p$. This is not surprising since X takes the values 0 and 1 and hence $X^n = X$. △

Example 7.6 (Moments of the exponential distribution)**.** Let $X \sim \text{Exp}(\lambda)$. From Example 7.3 its m.g.f. is

$$M(t) = \begin{cases} \dfrac{\lambda}{\lambda - t} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda. \end{cases}$$

Since $\lambda > 0$, we can differentiate around the origin and find

$$M'(t) = \lambda(\lambda - t)^{-2}, \quad M''(t) = 2\lambda(\lambda - t)^{-3}, \dots, \quad M^{(n)}(t) = n!\lambda(\lambda - t)^{-n-1}.$$

From this $E(X^n) = M^{(n)}(0) = n!\lambda^{-n}$ for positive integers n. △

For the next example, we recall that the general form of the Taylor expansion of a function f around the origin is

$$f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} t^n. \tag{7.2}$$

Example 7.7 (Moments of the standard normal)**.** Let $Z \sim \mathcal{N}(0, 1)$. From Example 7.2 we have $M_Z(t) = e^{t^2/2}$. Instead of differentiating this repeatedly, we can find the derivatives directly from the Taylor expansion.

Using the series for the exponential function we get

$$M_Z(t) = e^{t^2/2} = \sum_{k=0}^{\infty} \frac{\left(\frac{1}{2}t^2\right)^k}{k!} = \sum_{k=0}^{\infty} \frac{1}{2^k k!} t^{2k}. \tag{7.3}$$

Comparing this with

$$M_Z(t) = \sum_{n=0}^{\infty} \frac{M_Z^{(n)}(0)}{n!} t^n,$$

we can identify the values $M_Z^{(n)}(0)$ by matching the coefficients of the powers of t. Since only even powers t^{2k} appear in (7.3), we conclude that the coefficients of odd powers are zero, while

$$\frac{M_Z^{(2k)}(0)}{(2k)!} = \frac{1}{2^k k!}.$$

This gives us

$$E(Z^n) = M_Z^{(n)}(0) = \begin{cases} 0, & \text{if } n \text{ is odd,} \\ \frac{(2k)!}{2^k k!}, & \text{if } n = 2k \text{ is even.} \end{cases}$$

The expression for even moments simplifies by splitting $(2k)!$ into even and odd factors:

$$(2k)! = \left(\prod_{i=1}^{k} (2i) \right) \cdot \left(\prod_{j=1}^{k} (2j - 1) \right) = 2^k k! \cdot (2k - 1)!!$$

where the *double factorial* is

$$n!! = \begin{cases} n(n-2)(n-4) \cdots 1, & n > 0 \text{ is odd} \\ n(n-2)(n-4) \cdots 2, & n > 0 \text{ is even.} \end{cases}$$

We can summarize the moments of Z as follows:

$$E(Z^n) = \begin{cases} 0, & \text{if } n \text{ is odd,} \\ (n-1)!!, & \text{if } n \text{ is even.} \end{cases}$$

The fact that the odd moments are zero also follows from the fact that the density function $\varphi(x)$ of Z is an even function. △

Identification of distributions with moment generating functions.

Theorem 7.8. *Let X and Y be two random variables with moment generating functions $M_X(t) = E(e^{tX})$ and $M_Y(t) = E(e^{tY})$. Suppose there exists $\delta > 0$ such that for $t \in (-\delta, \delta)$, $M_X(t) = M_Y(t)$ and these are finite numbers. Then X and Y are equal in distribution.*

In the case when the random variable X takes only a finite number of values the m.g.f. takes a particularly simple form:

$$M_X(t) = E(e^{tX}) = \sum_k e^{kt} P(X = k). \tag{7.4}$$

When the moment generating function is in this form, we can read off directly the values of X and their probabilities.

Example 7.9. Suppose that X has moment generating function

$$M_X(t) = \tfrac{1}{5} e^{-17t} + \tfrac{1}{4} + \tfrac{11}{20} e^{2t}. \tag{7.5}$$

Then the possible values of X are $\{-17, 0, 2\}$ and

$$P(X = -17) = \tfrac{1}{5}, \quad P(X = 0) = \tfrac{1}{4}, \quad P(X = 2) = \tfrac{11}{20}. \tag{7.6}$$

Let us emphasize the logic. X must have the distribution given in (7.6) because (i) this probability mass function gives the moment generating function (7.5) and (ii) by Theorem 7.8 no other distribution can give this same moment generating function. △

Example 7.10. Find the distribution of Y if its moment generating function is

$$M_Y(t) = e^{17(e^t - 1)}.$$

We recognize the moment generating function as that of a Poisson with parameter 17. △

Moment generating function of a sum of independent random variables.

Theorem 7.11. *Suppose that X and Y are independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then for all real numbers t,*

$$M_{X+Y}(t) = M_X(t) M_Y(t). \tag{7.7}$$

Proof. The proof follows from the independence and Theorem 4.25:

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX} e^{tY}] = E[e^{tX}] E[e^{tY}] = M_X(t) M_Y(t). \qquad \square$$

Theorem 7.11 extends readily to sums of arbitrarily many random variables: if $X_1, X_2, \ldots, X_n$ are independent with sum $S = X_1 + X_2 + \cdots + X_n$, then

$$M_S(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t). \tag{7.8}$$

Example 7.12. Let $X_1, X_2, \dots, X_{20}$ be i.i.d. random variables with probability mass function $P(X_i = 2) = \frac{1}{3}$ and $P(X_i = 5) = \frac{2}{3}$. Let $S = X_1 + \cdots + X_{20}$. Find the moment generating function of S.

For each X_i we have $M_{X_i}(t) = \frac{1}{3}e^{2t} + \frac{2}{3}e^{5t}$. By (7.8), $M_S(t) = (\frac{1}{3}e^{2t} + \frac{2}{3}e^{5t})^{20}$. △

Together, Theorems 7.8 and 7.11 can be used to identify the distribution of a sum $X + Y$ of independent random variables, if we can identify M_{X+Y} as the moment generating function of a known distribution. This gives an alternative to calculating convolutions of probability mass functions or probability density functions.

Example 7.13 (Convolution of Poisson random variables revisited)**.** Suppose that $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$ and these are independent.

By Example 7.1 we have $M_X(t) = e^{\lambda(e^t-1)}$ and $M_Y(t) = e^{\mu(e^t-1)}$. Then $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)}e^{\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$. But this is the same as the moment generating function of a Poisson$(\lambda + \mu)$ random variable, hence $X + Y \sim \text{Poisson}(\lambda + \mu)$. △

Example 7.14 (Convolution of normal random variables)**.** Suppose that $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and these are independent.

By Example 7.2 we have $M_X(t) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2}$ and $M_Y(t) = e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2}$. Thus,

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2} = e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}.$$

This shows that $X+Y$ has the moment generating function of a $\mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ random variable. Consequently $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. △

The example above generalizes to a sum of any number of independent normal random variables. We combine it with Theorem 6.14 on the affine transformations of normal random variables to state the following theorem.

Theorem 7.15. *Assume $X_1, X_2, \dots, X_n$ are independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $a_i \neq 0$, and $b \in \mathbb{R}$. Let $X = a_1X_1 + \cdots + a_nX_n + b$. Then $X \sim \mathcal{N}(\mu, \sigma^2)$ where*

$$\mu = a_1\mu_1 + \cdots + a_n\mu_n + b \quad \textit{and} \quad \sigma^2 = a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2.$$

Example 7.16. Let $X \sim \mathcal{N}(1, 3)$ and $Y \sim \mathcal{N}(0, 4)$ be independent and $W = \frac{1}{2}X - Y + 6$. Identify the distribution of W.

According to Theorem 7.15, W has normal distribution with mean

$$E[W] = \tfrac{1}{2}E[X] - E[Y] + 6 = \tfrac{1}{2} - 0 + 6 = \tfrac{13}{2}$$

and variance

$$\text{Var}(W) = \tfrac{1}{4}\,\text{Var}(X) + \text{Var}(Y) = \tfrac{3}{4} + 4 = \tfrac{19}{4}.$$

Note that all terms in the variance calculation come with a plus sign, even though one random variable comes with a minus sign in the expression for W. △

The last example of this part shows how the moment generating function can be used as an alternative to convolution to derive the distribution of a sum of independent discrete random variables.

Example 7.17. Let X and Y be independent with probability mass functions

$$p_X(1) = \tfrac{1}{3}, \quad p_X(2) = \tfrac{1}{4}, \quad p_X(3) = \tfrac{1}{6}, \quad p_X(4) = \tfrac{1}{4}$$

and

$$p_Y(1) = \tfrac{1}{2}, \quad p_Y(2) = \tfrac{1}{3}, \quad p_Y(3) = \tfrac{1}{6}.$$

Find the probability mass function of $X + Y$.

We start by computing the moment generating functions of X and Y:

$$M_X(t) = \tfrac{1}{3}e^t + \tfrac{1}{4}e^{2t} + \tfrac{1}{6}e^{3t} + \tfrac{1}{4}e^{4t} \quad \text{and} \quad M_Y(t) = \tfrac{1}{2}e^t + \tfrac{1}{3}e^{2t} + \tfrac{1}{6}e^{3t}.$$

Then $M_{X+Y}(t) = M_X(t)M_Y(t)$ gives

$$\begin{aligned} M_{X+Y}(t) &= (\tfrac{1}{3}e^t + \tfrac{1}{4}e^{2t} + \tfrac{1}{6}e^{3t} + \tfrac{1}{4}e^{4t})(\tfrac{1}{2}e^t + \tfrac{1}{3}e^{2t} + \tfrac{1}{6}e^{3t}) \\ &= \tfrac{1}{6}e^{2t} + \tfrac{17}{72}e^{3t} + \tfrac{2}{9}e^{4t} + \tfrac{2}{9}e^{5t} + \tfrac{1}{9}e^{6t} + \tfrac{1}{24}e^{7t} \end{aligned}$$

where we expanded the product. By reading off the coefficients of e^{kt} for $k = 2, 3, \dots, 7$ we get the probability mass function of $X + Y$:

$$\begin{aligned} &p_{X+Y}(2) = \tfrac{1}{6}, \quad p_{X+Y}(3) = \tfrac{17}{72}, \quad p_{X+Y}(4) = \tfrac{2}{9}, \\ &p_{X+Y}(5) = \tfrac{2}{9}, \quad p_{X+Y}(6) = \tfrac{1}{9}, \quad p_{X+Y}(7) = \tfrac{1}{24}. \end{aligned}$$

△

Sketch of the proof of the central limit theorem.

We give a partial proof of the CLT by showing that the moment generating function of the standardized sum $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the moment generating function of the standard normal distribution. The convergence of moment generating functions implies the convergence of distributions. This is the content of the next theorem, which we state without proof.

Theorem 7.18 (Continuity theorem for moment generating functions)**.** *Suppose X is a random variable whose moment generating function $M_X(t)$ is finite in an interval $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Assume further that the moment generating functions of the random variables $Y_1, Y_2, Y_3, \dots$ satisfy*

$$\lim_{n\to\infty} M_{Y_n}(t) = M_X(t)$$

for all t in the interval $(-\varepsilon, \varepsilon)$. Then Y_n converges in distribution to X.

Specializing Theorem 7.18 to a standard normal X gives the following corollary.

Theorem 7.19. *Assume that the moment generating functions of the random variables $Y_1, Y_2, Y_3, \dots$ satisfy*

$$\lim_{n\to\infty} M_{Y_n}(t) = e^{\frac{t^2}{2}} \tag{7.9}$$

for all t in an interval $(-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Then for any $a \in \mathbb{R}$

$$\lim_{n\to\infty} P(Y_n \le a) = \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy.$$

We apply Theorem 7.19 to give a rough sketch of the proof of the CLT for the case where random variables have a finite moment generating function. The reader should appreciate how easy this is, compared to potentially trying to show that the distribution function of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to Φ.

Sketch of the proof of Theorem 6.17. We assume that $M_{X_i}(t)$ is finite and take $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. We prove limit (7.9) from which the conclusion (6.10) follows.

From the definition of the moment generating function and the fact that the random variables $X_1, \dots, X_n$ are independent,

$$M_{Y_n}(t) = E\big[e^{tY_n}\big] = E\Big[e^{t\frac{S_n - n\mu}{\sigma\sqrt{n}}}\Big] = E\Big[\exp\Big(\frac{t}{\sigma\sqrt{n}}\sum_{k=1}^{n}(X_k - \mu)\Big)\Big]$$

$$= E\Big[\prod_{k=1}^{n} e^{\frac{t}{\sigma\sqrt{n}}(X_k-\mu)}\Big] = \prod_{k=1}^{n} E\big[e^{\frac{t}{\sigma\sqrt{n}}(X_k-\mu)}\big]. \tag{7.10}$$

Inside each expectation we insert the second order Taylor approximation $e^x \approx 1 + x + x^2/2$ which is valid near 0 and then use additivity of expectation:

$$\begin{aligned} E\big[e^{\frac{t}{\sigma\sqrt{n}}(X_k-\mu)}\big] &\approx E\Big[1 \;+\; \frac{t}{\sigma\sqrt{n}}(X_k-\mu) \;+\; \frac{t^2}{2\sigma^2 n}(X_k-\mu)^2\Big] \\ &= 1 \;+\; \frac{t}{\sigma\sqrt{n}}E[X_k-\mu] \;+\; \frac{t^2}{2\sigma^2 n}E\big[(X_k-\mu)^2\big] \\ &= 1 \;+\; \frac{t}{\sigma\sqrt{n}}\cdot 0 \;+\; \frac{t^2}{2\sigma^2 n}\cdot\sigma^2 \\ &= 1 \;+\; \frac{t^2}{2n}. \end{aligned}$$

Put this back above in (7.10) to get

$$M_{Y_n}(t) \approx \Big(1 + \frac{t^2}{2n}\Big)^n \longrightarrow e^{t^2/2} \qquad \text{as } n \to \infty.$$

We have verified (7.9).

Making this proof watertight requires error bounds for the $\approx$ steps. Note that as $n \to \infty$ the Taylor approximation gains accuracy because the quantity $\frac{t}{\sigma\sqrt{n}}(X_k - \mu)$ is converging to zero. □

With the characteristic function the CLT can be proved with an argument similar to the one above, without any additional assumptions beyond finite mean and variance.

Exercises

Exercise 7.1.

(a) Let $Z \sim \text{Unif}[0,1]$. Find the moment generating function $M_Z(t)$ of Z.

(b) For each positive integer n, let X_n be a *uniform* random variable on the set $\{\frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1\}$, by which we mean that $P(X_n = \frac{k}{n}) = \frac{1}{n}$ for each $k \in \{1, 2, \dots, n\}$. Use Theorem 7.18 to prove a limit in distribution $X_n \xrightarrow{d} X$ and identify the limit.

Exercise 7.2. Recall the definition of the gamma distribution: for real $r, \lambda > 0$, $X \sim \text{Gamma}(r, \lambda)$ means that X has probability density function

$$f_X(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} \qquad \text{for } x > 0,$$

and $f_X(x) = 0$ for $x \leq 0$, where the gamma function is defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx \qquad \text{for } r > 0.$$

(a) For all parameter values $r, \lambda > 0$, find the moment generating function of $X \sim \text{Gamma}(r, \lambda)$.

(b) For $X \sim \text{Gamma}(r, \lambda)$, calculate all the moments $E(X^n)$ for positive integers n and the variance.

(c) If X and Y are independent with distributions $X \sim \text{Gamma}(r, \lambda)$ and $Y \sim \text{Gamma}(s, \lambda)$, find the distribution of $X + Y$.

(d) Use the moment generating function of the exponential distribution to determine the distribution of a sum of n i.i.d. $\text{Exp}(\lambda)$ random variables.

Exercise 7.3. Let $X \sim \text{Geom}(p)$.

(a) Compute the moment generating function $M_X(t)$ of X.

(b) Use the moment generating function to compute the mean and the variance of X.

Exercise 7.4. Let $Z \sim \mathcal{N}(0, 1)$ and $Y = e^Z$. Y is called a *log-normal* random variable.

(a) Find the probability density function of Y.

(b) Find the nth moment $E(Y^n)$ of Y.
Hint. Instead of trying to compute the moment generating function of Y, relate the nth moment of Y to an expectation of Z.

Exercise 7.5. Fix $0 < \lambda < \infty$ and for integers $n > \lambda$ let $S_n \sim \text{Bin}(n, \lambda/n)$ and $Y \sim \text{Poisson}(\lambda)$. Prove the weak limit $S_n \xrightarrow{d} Y$ with moment generating functions.

Chapter 8

Conditional expectation

8.1. Conditional distributions

Expectations of random variables are calculated with probability mass functions and probability density functions. Conditional expectations of random variables are calculated with conditional probability mass functions and conditional probability density functions. We define these first for cases where a discrete random variables is conditioned on the value of another discrete random variable, and then for two jointly absolutely continuous random variables.

Definition 8.1. Let X and Y be discrete random variables. Let $y \in \mathbb{R}$ be point such that $P(Y = y) > 0$. Then the **conditional probability mass function of** X **given** $Y = y$ is the function $p_{X|Y}(x \,|\, y)$ of possible values x of X, defined as follows:

$$p_{X|Y}(x \,|\, y) = P(X = x \,|\, Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \tag{8.1}$$

The **conditional expectation of** X **given** $Y = y$ is

$$E[X|Y = y] = \sum_x x\, p_{X|Y}(x \,|\, y). \tag{8.2}$$

△

For a fixed y, as a function of x the conditional probability mass function $p_{X|Y}(x|y)$ behaves like a regular probability mass function: its values are nonnegative and sum up to one:

$$\sum_x p_{X|Y}(x|y) = \frac{1}{P(Y = y)} \sum_x P(X = x, Y = y) = \frac{P(Y = y)}{P(Y = y)} = 1. \tag{8.3}$$

Example 8.2 (Multinomial)**.** Consider n independent repetitions of a trial with three possible outcomes labeled $1, 2$, and 3. In each trial these outcomes appear with probabilities p_1, p_2, and p_3, respectively, where $p_1 + p_2 + p_3 = 1$. Let X_i be the number of times outcome i appears among the n trials. Find the conditional probability mass function $p_{X_2|X_1}(\ell \,|\, m)$ and the conditional expectation $E[X_2 \,|\, X_1 = m]$. Before the calculation you should try to guess the answer from intuition.

Let $0 \le m \le n$. Since $X_1 + X_2 + X_3 = n$ and from Example 3.41 the marginal distribution of X_1 is $\text{Bin}(n, p_1)$, we deduce that for $0 \le \ell \le n - m$,

$$\begin{aligned} p_{X_2|X_1}(\ell \,|\, m) = P(X_2 = \ell \,|\, X_1 = m) &= \frac{P(X_1 = m,\, X_2 = \ell)}{P(X_1 = m)} \\ &= \frac{P(X_1 = m,\, X_2 = \ell,\, X_3 = n - m - \ell)}{P(X_1 = m)} \\ &= \frac{\binom{n}{m,\,\ell,\,n-m-\ell}\, p_1^m\, p_2^\ell\, p_3^{n-m-\ell}}{\binom{n}{m}\, p_1^m\, (p_2 + p_3)^{n-m}} \\ &= \frac{(n-m)!}{\ell!(n-m-\ell)!} \left(\frac{p_2}{p_2+p_3}\right)^{\ell} \left(\frac{p_3}{p_2+p_3}\right)^{n-m-\ell}. \end{aligned}$$

In the third line we used the joint probability mass function for the multinomial distribution. The formula tells us that, given $X_1 = m$, $X_2 \sim \text{Bin}(n - m, \frac{p_2}{p_2+p_3})$. Consequently, by the formula for the mean of a binomial,

$$E[X_2 \,|\, X_1 = m] = (n - m)\frac{p_2}{p_2 + p_3}.$$

△

Next the definition in the jointly continuous case.

Definition 8.3. Let X and Y be jointly continuous random variables with joint density function $f_{X,Y}(x, y)$. For those $y \in \mathbb{R}$ such that $f_Y(y) > 0$ we make the following definitions.

The **conditional density function of X, given $Y = y$,** is denoted by $f_{X|Y}(x \,|\, y)$ and defined as

$$f_{X|Y}(x \,|\, y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \tag{8.4}$$

The **conditional probability that $X \in A$, given $Y = y$,** is

$$P(X \in A \,|\, Y = y) = \int_A f_{X|Y}(x \,|\, y)\, dx. \tag{8.5}$$

The **conditional expectation of X, given $Y = y$,** is

$$E[X \,|\, Y = y] = \int_{-\infty}^{\infty} x\, f_{X|Y}(x \,|\, y)\, dx. \tag{8.6}$$

△

We check that the conditional density function integrates to 1, as any legitimate density function should. Suppose that $f_Y(y) > 0$ and recall that $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(w, y)dw$. Then

$$\int_{-\infty}^{\infty} f_{X|Y}(x \,|\, y)\, dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(w, y)dw}\, dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x, y)dx}{\int_{-\infty}^{\infty} f_{X,Y}(w, y)dw} = 1.$$

Example 8.4. Let (X, Y) be a uniformly chosen random point on a disk D centered at $(0, 0)$ with radius r_0. From Example 2.42 we recall the joint density function

$$f_{X,Y}(x, y) = \begin{cases} \dfrac{1}{\pi r_0^2} & \text{if } (x, y) \in D \\ 0 & \text{if } (x, y) \notin D, \end{cases}$$

and the marginal density functions for $-r_0 < x < r_0$ and $-r_0 < y < r_0$

$$f_X(x) = \frac{2\sqrt{r_0^2 - x^2}}{\pi r_0^2} \quad \text{and} \quad f_Y(y) = \frac{2\sqrt{r_0^2 - y^2}}{\pi r_0^2}.$$

Formula (8.4) for the conditional density function of X, given $Y = y$, gives

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{2\sqrt{r_0^2 - y^2}}.$$

To determine the domain, first we need $-r_0 < y < r_0$ to guarantee that $f_Y(y) > 0$. Second, to have $f_{X,Y}(x, y) > 0$ we need $(x, y) \in D$ which is equivalent to $-\sqrt{r_0^2 - y^2} < x < \sqrt{r_0^2 - y^2}$.

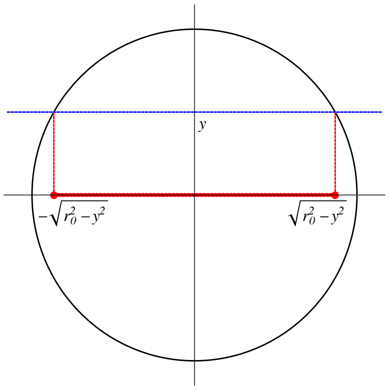

Figure 1. Conditional on $Y = y$, the point (X, Y) lies on the intersection of the blue dashed line with the disk, and the conditional distribution of X is uniform on the solid red line on the x-axis.

So the complete answer is that for $-r_0 < y < r_0$,

$$f_{X|Y}(x\,|\,y) = \frac{1}{2\sqrt{r_0^2 - y^2}} \qquad \text{for } -\sqrt{r_0^2 - y^2} < x < \sqrt{r_0^2 - y^2},$$

and zero otherwise. In other words, given $Y = y \in (-r_0, r_0)$, X is uniformly distributed on the interval $(-\sqrt{r_0^2 - y^2}\,, \sqrt{r_0^2 - y^2}\,)$. △

Example 8.5. Let X and Y be independent exponential random variables with parameter μ and $Z = X + Y$. Find the conditional density function of X given $Z = z$ and the conditional expectation $E[X|Z = z]$.

A word problem formulation of this question could go as follows. Times between successive calls to the tech support line are independent exponential random variables with parameter μ. Given that the second call of the day came at time z, what is the distribution of the time of the first call?

To apply the formula $f_{X|Z}(x|z) = \frac{f_{X,Z}(x,z)}{f_Z(z)}$ we need the joint density function of (X, Z) and the marginal density function of Z. Example 3.47 deduced the gamma density $f_Z(z) = \mu^2 z e^{-\mu z}$ for $z > 0$. We give two approaches for finding $f_{X,Z}$.

(i) Find the joint cumulative distribution function $F_{X,Z}(x, z)$ and apply (2.41). Since $X, Y > 0$ with probability 1, it is enough to consider values $0 < x < z$ for (X, Z). Utilize the independence of X and Y below.

$$\begin{aligned} F_{X,Z}(x,z) &= P(X \le x, Z \le z) = P(X \le x, X + Y \le z) \\ &= \iint\limits_{s \le x,\, s+t \le z} f_{X,Y}(s,t)\, ds\, dt = \iint\limits_{s \le x,\, t \le z-s} f_X(s) f_Y(t)\, ds\, dt \\ &= \int_0^x \int_0^{z-s} \mu^2 e^{-\mu(t+s)}\, dt\, ds = 1 - e^{-\mu x} - \mu x e^{-\mu z}. \end{aligned}$$

By (2.41)

$$f_{X,Z}(x,z) = \frac{\partial^2}{\partial x \partial z} F_{X,Z}(x,z) = \mu^2 e^{-\mu z} \qquad \text{for } 0 < x < z.$$

For other values we can take $f_{X,Z}(x, z) = 0$.

(ii) Here is a quick argument that requires very little computation that identifies the joint density function $f_{X,Z}$. We begin with the expectation formula

$$E[g(X,Z)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,z)\, f_{X,Z}(x,z)\, dz\, dx, \tag{8.7}$$

where g is an arbitrary function. Compute the left-hand side with the joint density of (X, Y):

$$\begin{aligned} E[g(X,Z)] = E[g(X, X+Y)] &= \int_0^{\infty} \int_0^{\infty} g(x, x+y)\, f_X(x)\, f_Y(y)\, dy\, dx \\ &= \int_0^{\infty} \left(\int_0^{\infty} g(x, x+y)\, \mu^2 e^{-\mu(x+y)}\, dy \right) dx \\ &= \int_0^{\infty} \left(\int_x^{\infty} g(x,z)\, \mu^2 e^{-\mu z}\, dz \right) dx \\ &= \iint_{0<x<z} g(x,z)\, \mu^2 e^{-\mu z}\, dz\, dx. \end{aligned}$$

The variable in the inner integral changed from y to $z = x + y$, for a fixed x. Comparison of the last representation of $E[g(X, Z)]$ with (8.7) identifies the joint density function as $f_{X,Z}(x, z) = \mu^2 e^{-\mu z}$ for $0 < x < z$ and zero elsewhere.

The conditional density function can now be written down:

$$f_{X|Z}(x|z) = \frac{\mu^2 e^{-\mu z}}{\mu^2 z e^{-\mu z}} = \frac{1}{z} \qquad \text{for } 0 < x < z.$$

This shows that, given $Z = z$, the conditional distribution of X is uniform on the interval $(0, z)$. Thus the conditional expectation is $E[X|Z = z] = \frac{z}{2}$. △

Everything we know about ordinary unconditioned expectations is valid for conditional expectations, with the right alterations in the statements and formulas.

The conditional expectation formulas extend to expectations of functions of X exactly as in Theorem 4.12 for unconditioned expectations.

Theorem 8.6. *For the conditional expectation of $g(X)$ given that $Y = y$ we have the following formulas provided the expectations are well-defined.*

(i) *In the discrete case*

$$E[g(X)\,|\,Y=y] = \sum_x g(x)\,p_{X|Y}(x\,|\,y) \tag{8.8}$$

for y such that $P(Y = y) > 0$.

(ii) *In the jointly continuous case*

$$E[g(X)\,|\,Y=y] = \int_{-\infty}^{\infty} g(x)\,f_{X|Y}(x\,|\,y)\,dx. \tag{8.9}$$

for y such that $f_Y(y) > 0$.

In addition to representing conditioned quantities, conditional expectations turn out to be useful for computing unconditioned expectations. This is based on the averaging identities collected in the next theorem. The message of the theorem is that unconditioned expectations can be calculated by averaging conditional expectations.

Theorem 8.7. *We have the following averaging identities. Assume that the expectations of $g(X)$ below are well-defined.*

(i) *In the discrete case*

$$p_X(x) = \sum_y p_{X|Y}(x\,|\,y)\,p_Y(y) \tag{8.10}$$

and

$$E[g(X)] = \sum_y E[g(X)\,|\,Y=y]\,p_Y(y). \tag{8.11}$$

(ii) *In the jointly continuous case*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)\,f_Y(y)\,dy \tag{8.12}$$

and

$$E[g(X)] = \int_{-\infty}^{\infty} E[g(X)\,|\,Y=y]\,f_Y(y)\,dy. \tag{8.13}$$

These identities reveal why there is no harm in restricting the definitions of $p_{X|Y}(x\,|\,y)$, $f_{X|Y}(x\,|\,y)$ and $E[g(X)\,|\,Y=y]$ to those points y for which $p_Y(y) > 0$ or $f_Y(y) > 0$, whichever the case may be. Namely, in the sums in (8.10)–(8.11) contributions come only from those y for which $p_Y(y) > 0$. In the integrals in (8.12)–(8.13) contributions come only from those y for which $f_Y(y) > 0$.

Proof. We leave the proofs of the discrete cases as Exercise 8.2. Rearranging the terms in (8.4) gives the identity

$$f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)\,f_Y(y). \tag{8.14}$$

Integration over y gives

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)\,f_Y(y)\,dy,$$

and (8.12) has been verified.

We deduce (8.13) with a string of equalities that utilize the definitions and identity (8.14).

$$\begin{aligned}\int_{-\infty}^{\infty} E[g(X)|Y=y]\,f_Y(y)\,dy &= \int_{-\infty}^{\infty}\biggl(\int_{-\infty}^{\infty} g(x) f_{X|Y}(x\,|\,y)\,dx\biggr) f_Y(y)\,dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\,f_{X|Y}(x\,|\,y)\,f_Y(y)\,dx\,dy \\ &= \int_{-\infty}^{\infty}\int_{\infty}^{\infty} g(x)\,f_{X,Y}(x,y)\,dx\,dy \;=\; E[g(X)]. \qquad \square\end{aligned}$$

Remark 8.8 (Conditional probability mass function, given an event)**.** Definition (8.1) above includes as a special case conditioning on an event B. We would naturally define the conditional probability mass function $p_{X|B}$ of X, given B, by

$$p_{X|B}(k) = P(X=k\,|\,B) = \frac{P(\{X=k\}\cap B)}{P(B)}.$$

If I_B denotes the indicator random variable of B, then this is the same as

$$p_{X|I_B}(k\,|\,1) = P(X=k\,|\,I_B=1) = P(X=k\,|\,B) = p_{X|B}(k).$$

△

The next example illustrates a situation where the conditional information, given an event, is naturally available.

Example 8.9. Let X denote the number of customers that arrive in my store tomorrow. If the day is rainy X is Poisson(λ), and if the day is dry X is Poisson(μ). Suppose that the probability it rains tomorrow is 10%. Find the probability mass function and expectation of X.

Let B be the event that it rains tomorrow. We have $P(B) = 0.1$. The conditional probability mass functions and conditional expectations for tomorrow's arrivals are

$$\begin{gathered} p_{X|B}(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad p_{X|B^c}(k) = e^{-\mu}\frac{\mu^k}{k!}, \\ E(X\,|\,B) = \lambda, \qquad \text{and} \qquad E(X\,|\,B^c) = \mu. \end{gathered}$$

From (8.10) the unconditional probability mass function is

$$p_X(k) = P(B)\,p_{X|B}(k) + P(B^c)\,p_{X|B^c}(k) = \frac{1}{10}\cdot e^{-\lambda}\frac{\lambda^k}{k!} + \frac{9}{10}\cdot e^{-\mu}\frac{\mu^k}{k!}.$$

From (8.11) the expected number of customers is

$$E[X] = P(B)E(X\,|\,B) + P(B^c)E(X\,|\,B^c) = \frac{\lambda}{10} + \frac{9\mu}{10}.$$

△

Example 8.10. Suppose that X and Y are independent Poisson random variables with parameters λ and μ and let $Z = X+Y$. Find the conditional probability mass function of X and conditional expectation of X, given $Z = \ell$. Verify the averaging identity (8.10) for p_X.

The joint probability mass function of X, Z comes by taking advantage of the independence of X and Y. Since $0 \le X \le Z$, we only need to consider $0 \le k \le \ell$ in the probability below:

$$\begin{aligned} P(X = k, Z = \ell) &= P(X = k, X + Y = \ell) = P(X = k, Y = \ell - k) \\ &= P(X = k)P(Y = \ell - k) = \tfrac{\lambda^k}{k!}e^{-\lambda}\tfrac{\mu^{\ell-k}}{(\ell-k)!}e^{-\mu}. \end{aligned}$$

Recall from Example 3.44 or Example 7.13 that $Z = X + Y \sim \text{Poisson}(\lambda + \mu)$. Now we have the information needed for deducing the conditional probability mass function of X. For $0 \le k \le \ell$,

$$\begin{aligned} p_{X|Z}(k|\ell) &= \frac{p_{X,Z}(k,\ell)}{p_Z(\ell)} = \frac{\frac{\lambda^k}{k!}e^{-\lambda}\frac{\mu^{\ell-k}}{(\ell-k)!}e^{-\mu}}{\frac{(\lambda+\mu)^\ell}{\ell!}e^{-(\lambda+\mu)}} = \frac{\lambda^k\mu^{\ell-k}}{(\lambda+\mu)^\ell}\frac{\ell!}{k!(\ell-k)!} \\ &= \binom{\ell}{k}\left(\tfrac{\lambda}{\lambda+\mu}\right)^k\left(\tfrac{\mu}{\lambda+\mu}\right)^{\ell-k}. \end{aligned}$$

This says that, given $Z = \ell$, the conditional distribution of X is $\text{Bin}(\ell, \frac{\lambda}{\lambda+\mu})$. Recalling that the mean of $\text{Bin}(n, p)$ equals np, we can write down the conditional expectations without further computation: $E[X|Z = \ell] = \ell\frac{\lambda}{\lambda+\mu}$.

We check that the averaging identity (8.10) gives back the probability mass function of X:

$$\begin{aligned} \sum_\ell p_{X|Z}(k|\ell)\, p_Z(\ell) &= \sum_{\ell=k}^{\infty}\binom{\ell}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{\ell-k}\frac{(\lambda+\mu)^\ell}{\ell!}e^{-(\lambda+\mu)} \\ &= \frac{\lambda^k e^{-(\lambda+\mu)}}{k!}\sum_{\ell=k}^{\infty}\frac{\mu^{\ell-k}}{(\ell-k)!} = \frac{\lambda^k e^{-(\lambda+\mu)}}{k!}\sum_{j=0}^{\infty}\frac{\mu^j}{j!} = \frac{\lambda^k e^{-(\lambda+\mu)}}{k!}\cdot e^{\mu} \\ &= e^{-\lambda}\frac{\lambda^k}{k!}. \end{aligned}$$

On the last line we recognize the Poisson(λ) probability mass function $p_X(k)$.

Here is a story to illuminate the example. Count cars and trucks passing an intersection during a fixed time interval. Let X be the number of cars, Y the number of trucks, and Z their total. Assume that X and Y are independent Poisson random variables. The calculation above can be interpreted as follows. Given that there were $Z = \ell$ total vehicles, the distribution of the random number X of cars is the same as that obtained by marking each vehicle independently as a car with probability $\frac{\lambda}{\lambda+\mu}$. This is a special feature of the Poisson distribution. △

Constructing joint probability distributions. Rearranging (8.1) expresses the joint probability mass function in terms of the conditional and marginal probability mass functions:

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)\, p_Y(y). \tag{8.15}$$

Even though (8.1) was valid only when $p_Y(y) > 0$, this is not an issue now because if $p_Y(y) = 0$ then $p_{X,Y}(x, y) = 0$ as well.

Identity (8.15) can be used to define a joint probability mass function when a marginal and a conditional probability mass function are given. The next example illustrates this idea and the averaging identity (8.11).

Example 8.11. Suppose n people apply for a job. Each applicant has to pass two tests. Each person independently passes the first test with probability p. Only those who pass the first test can take the second test. Each person independently passes the second test with probability r. Let M be the number of people who pass the first test and L the number of people who pass the second test after passing the first one. Find the joint probability mass function of M, L and the mean of L.

The information tells us that $M \sim \text{Bin}(n,p)$, and given that $M = m$, $L \sim \text{Bin}(m, r)$. In terms of probability mass functions, for $0 \le \ell \le m \le n$,

$$p_M(m) = \binom{n}{m} p^m (1-p)^{n-m} \quad \text{and} \quad p_{L|M}(\ell|m) = \binom{m}{\ell} r^\ell (1-r)^{m-\ell}.$$

Furthermore, from knowing the mean of a binomial, we deduce $E[L|M = m] = mr$.

By (8.15) the joint probability mass function is, for $0 \le \ell \le m \le n$,

$$p_{M,L}(m,\ell) = p_{L|M}(\ell|m)\, p_M(m) = \binom{m}{\ell} r^\ell (1-r)^{m-\ell} \cdot \binom{n}{m} p^m (1-p)^{n-m}.$$

By the averaging identity (8.11) and since the mean of a $\text{Bin}(n, p)$ is np,

$$E(L) = \sum_m E[L|M = m] p_M(m) = r \sum_{m=0}^{n} m\, p_M(m) = rE(M) = npr.$$

Exercise 8.3 asks you to find the marginal probability mass function p_L of L. △

In complete analogy with the discrete case, we can use the identity

$$f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y)$$

to define the joint density function $f_{X,Y}$ from a marginal and a conditional density function. The example below illustrates this and Exercise 8.4 continues the theme.

Example 8.12. Let Y be a standard normal random variable. Then let X be another normal random variable with variance 1 whose mean is the value Y just observed. Find the joint density function of (X, Y). Then, suppose we observe $X = x$. How is Y now distributed?

The problem statement gives these density functions for all real x, y:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \varphi(y) \quad \text{and} \quad f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi}} e^{-(x-y)^2/2} = \varphi(x-y)$$

where φ denotes the standard normal density function. The joint density function is given by

$$f_{X,Y}(x, y) = f_{X|Y}(x|y) f_Y(y) = \frac{1}{2\pi} e^{-\frac{1}{2}y^2 - \frac{1}{2}(x-y)^2} = \varphi(y)\varphi(x-y) \quad \text{for all real } x, y.$$

Once we observe $X = x$, the distribution of Y should be conditioned on $X = x$. First find the marginal density function of X.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} \varphi(y)\varphi(x-y)\,dy = \varphi * \varphi(x) = \frac{1}{\sqrt{4\pi}}e^{-x^2/4}.$$

We observed that the integral above is the convolution $\varphi*\varphi(x)$, the density function of the sum of two independent standard normals, which is the $\mathcal{N}(0,2)$ density function. This way we can identify f_X above without explicit integration.

From the definition of the conditional density function we have, for all real x and y,

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{\sqrt{\pi}}e^{-\frac{1}{2}y^2 - \frac{1}{2}(x-y)^2 + \frac{1}{4}x^2} = \frac{1}{\sqrt{\pi}}e^{-(y-\frac{x}{2})^2}.$$

The conclusion is that, given $X = x$, we have $Y \sim \mathcal{N}(\frac{x}{2}, \frac{1}{2})$. △

Marking Poisson arrivals. We close this section with a property of Poisson random variables that is important for applications. It is another application of conditioning as a step towards finding a joint distribution.

Example 8.13. Suppose the number of customers that arrive in my store during the course of a day is a Poisson(λ) random variable X. At the door each customer receives a randomly chosen coupon for a discount. The coupons come in three types labeled 1, 2 and 3. A coupon is of type i with probability p_i for $i = 1, 2, 3$, and naturally $p_1 + p_2 + p_3 = 1$. Assume that the coupons given to customers are independent. Let X_i be the number of customers who received a type i coupon, so that $X_1 + X_2 + X_3 = X$. Find the joint distribution of (X_1, X_2, X_3) and check whether they are independent or not.

First we need to understand the nature of the experiment. Suppose $X = n$ is given. Then the types of coupons come from n repeated independent trials, each with possible outcomes $1, 2,$ or 3, which have probabilities p_1, p_2, p_3. The random variable X_i is the number of trials that resulted in a type i outcome. Thus, given $X = n$, we have $(X_1, X_2, X_3) \sim \text{Mult}(n, 3, p_1, p_2, p_3)$.

Now let $k_1, k_2, k_3 \in \{0, 1, 2, \dots\}$ and set $k = k_1 + k_2 + k_3$.

$$\begin{aligned}
&P(X_1 = k_1, X_2 = k_2, X_3 = k_3) \\
&\quad = P(X_1 = k_1, X_2 = k_2, X_3 = k_3, X = k) \\
&\quad = P(X = k)\, P(X_1 = k_1, X_2 = k_2, X_3 = k_3 \,|\, X = k) \\
&\quad = \frac{e^{-\lambda}\lambda^k}{k!} \cdot \frac{k!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \\
&\quad = \frac{e^{-p_1\lambda}(p_1\lambda)^{k_1}}{k_1!} \cdot \frac{e^{-p_2\lambda}(p_2\lambda)^{k_2}}{k_2!} \cdot \frac{e^{-p_3\lambda}(p_3\lambda)^{k_3}}{k_3!}.
\end{aligned} \tag{8.16}$$

In the passage from line 3 to line 4 we used the *conditional joint probability mass function* of (X_1, X_2, X_3), given that $X = k$, namely

$$P(X_1 = k_1, X_2 = k_2, X_3 = k_3 \,|\, X = k) = \frac{k!}{k_1!k_2!k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3},$$

which came from the description of the problem. In the last equality of (8.16) we cancelled $k!$ and then used both $k = k_1 + k_2 + k_3$ and $p_1 + p_2 + p_3 = 1$.

We recognize Poisson probabilities on the last line of (8.16). To verify that the marginal distribution of X_1 is Poisson, sum away the other variables:

$$\begin{aligned}
P(X_1 = k) &= \sum_{k_2=0}^{\infty} \sum_{k_3=0}^{\infty} P(X_1 = k, X_2 = k_2, X_3 = k_3) \\
&= \frac{e^{-p_1\lambda}(p_1\lambda)^k}{k!} \cdot \sum_{k_2=0}^{\infty} \frac{e^{-p_2\lambda}(p_2\lambda)^{k_2}}{k_2!} \cdot \sum_{k_3=0}^{\infty} \frac{e^{-p_3\lambda}(p_3\lambda)^{k_3}}{k_3!} \\
&= \frac{e^{-p_1\lambda}(p_1\lambda)^k}{k!}.
\end{aligned}$$

This last calculation works the same for X_2 and X_3. Hence (8.16) can be completed to read

$$P(X_1 = k_1, X_2 = k_2, X_3 = k_3) = P(X_1 = k_1)P(X_2 = k_2)P(X_3 = k_3).$$

The conclusion is that X_1, X_2, X_3 are independent with Poisson marginals $X_i \sim \text{Poisson}(p_i\lambda)$.

By any rights, the conclusion that the Poisson distribution was preserved by the operation of handing out the coupons should surprise the reader. However, it seems downright counterintuitive that the random variables X_i are independent. Since $X_2 + X_3 = X - X_1$, should it not be the case that if X_1 is way above its mean $p_1\lambda$, we would expect X_2 and X_3 to be below their means? Or, going in the other direction, if $\lambda = 100$ we would expect around 100 customers on a given day. However, if we are told that 1000 type 1 coupons were given, we may think that there was a rush of customers, implying an increased number of type 2 and type 3 coupons. Yet neither of these “intuitive” explanations accounts for what really happens.

This process of cataloging each customer as a particular *type* is called *marking* in the probability literature. There was nothing special about splitting the Poisson number of customers into three categories. It works the same way for any number of labels attached to the arrivals. △

8.2. Conditional expectation

The conditional expectation formulas given in the previous section come from a general definition.

Definition 8.14. Let X and Y be two random variables defined on $(\Omega, \mathcal{F}, P)$. Assume the expectation EX is finite. Then there exists a function of the random variable Y that we denote temporarily by $v(Y)$ that satisfies the identity

$$E[v(Y)\,h(Y)] = E[X\,h(Y)] \tag{8.17}$$

for all bounded Borel functions h. The random variable $v(Y(\omega))$ is called the **conditional expectation of X, given Y** and denoted by $E(X \,|\, Y)(\omega)$. Furthermore, $E(X \,|\, Y)$ has finite expectation. △

The definition requires several points of explanation. We begin by clarifying the notation and terminology. $E(X \,|\, Y)$ is a single piece of notation that represents a particular function on Ω, in other words, a random variable, and $E(X \,|\, Y)(\omega)$ is the value of this random variable at the sample point $\omega \in \Omega$. The value of $E(X \,|\, Y)$ is determined by the value of Y through the connection $E(X \,|\, Y)(\omega) = v(Y(\omega))$.

The distinction between $E[X \,|\, Y = y]$ and $E(X \,|\, Y)$ is that the first one is a number, and the second one a random variable. They are differentiated by the terminology: $E[X|Y = y]$ is the *conditional expectation of* X *given* $Y = y$, while $E(X \,|\, Y)$ is the *conditional expectation of* X *given* Y.

Definition 8.14 asserts the existence of the random variable $E(X \,|\, Y)$ that is a function of Y and satisfies

$$E[E(X \,|\, Y)\, h(Y)] = E[X\, h(Y)] \tag{8.18}$$

for all bounded functions h. This needs proof, so the definition is really a definition and a theorem wrapped together. The proof requires advanced tools of measure theory so we omit it.

Another necessary aspect of the definition is that the object defined should be unique in some sense. The precise property is this: if $\widetilde{v}$ is another function such that

$$E[\widetilde{v}(Y)\, h(Y)] = E[X\, h(Y)] \tag{8.19}$$

for all bounded Borel functions h, then $P\{v(Y) = \widetilde{v}(Y)\} = 1$. In other words, the random variables $v(Y)$ and $\widetilde{v}(Y)$ are equal almost surely.

The class of functions h for which identity (8.18) is required can be reduced. For example, it is enough to have (8.18) for all indicator functions $h = I_B$ of Borel subsets of $\mathbb{R}$.

In the next two theorems we connect Definition 8.14 with the conditional expectations given $Y = y$ calculated in the previous section. The first theorem shows that the abstract conditional expectation $E[\,g(X) \,|\, Y]$ can be obtained from the concrete ones calculated previously.

Theorem 8.15. *Assume either that* (X, Y) *are discrete or that they have a joint density function* f. *Let* g *be a function such that* $g(X)$ *has a finite expectation. Then* $E[\,g(X) \,|\, Y](\omega) = v(Y(\omega))$ *for the function* $v(y) = E[\,g(X) \,|\, Y = y]$ *defined in Theorem 8.6.*

Proof. We prove the jointly continuous case and leave the discrete case as Exercise 8.6. In the continuous case, the function v is defined by (8.9) as

$$v(y) = \int_{-\infty}^{\infty} g(x)\, f_{X|Y}(x \,|\, y)\, dx \tag{8.20}$$

for those points y for which $f_Y(y) > 0$.

We check that (8.17) holds with this $v(Y)$, for the random variable $g(X)$ in place of X. Let h be an arbitrary bounded function. The calculation below uses the definitions of the joint, marginal and conditional density functions.

$$\begin{aligned}
E[v(Y)\,h(Y)] &= \int_{-\infty}^{\infty} v(y)\,h(y)\,f_Y(y)\,dy \\
&= \int_{-\infty}^{\infty} \biggl(\int_{-\infty}^{\infty} g(x)\,f_{X|Y}(x\,|\,y)\,dx \biggr) h(y)\,f_Y(y)\,dy \\
&= \iint_{\mathbb{R}^2} g(x)\,h(y)\,f_{X|Y}(x\,|\,y)\,f_Y(y)\,dx\,dy \\
&= \iint_{\mathbb{R}^2} g(x)\,h(y)\,f(x,y)\,dx\,dy = E[g(X)\,h(Y)]
\end{aligned} \tag{8.21}$$

Thus (8.17) has been verified and thereby we have established that $E[g(X)\,|\,Y] = v(Y)$ with v given in (8.20).

The calculation in (8.21) above suffered no harm from the fact that $v(y)$ was defined only when $f_Y(y) > 0$. If desired, we can assign arbitrary values to $v(y)$ when $f_Y(y) = 0$ without changing the value of the integral $\int_{-\infty}^{\infty} v(y)\,h(y)\,f_Y(y)\,dy$ on the first line. □

The next theorem shows that the abstract Definition 8.14 leads to the concrete formulas calculated in the previous section. However, we do not have enough integration theory at our disposal to prove this for the jointly continuous case. Hence we take care only of the discrete case.

Theorem 8.16. *Let X and Y be discrete random variables on $(\Omega, \mathcal{F}, P)$ and g a function such that $g(X)$ has finite expectation. Suppose the function v satisfies $E[v(Y)\,h(Y)] = E[g(X)\,h(Y)]$ for all bounded Borel functions h. Then*

$$v(y) = \sum_x g(x)\,p_{X|Y}(x|y) \qquad \text{for all } y \text{ such that } p_Y(y) > 0.$$

Proof. The assumption is that

$$\begin{aligned}
\sum_y v(y)h(y)p_Y(y) &= \sum_{x,y} g(x)h(y)p_{X,Y}(x,y) \\
&= \sum_y \biggl(\sum_x g(x)\,p_{X|Y}(x|y) \biggr) h(y)p_Y(y)
\end{aligned}$$

where the second equality came from writing $p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y)$. Now let y_0 be any value such that $p_Y(y_0) > 0$, and take h to be

$$h(y) = \begin{cases} \dfrac{1}{p_Y(y_0)}, & y = y_0 \\ 0, & y \neq y_0. \end{cases}$$

With this h the equality above becomes

$$v(y_0) = \sum_x g(x)\,p_{X|Y}(x|y_0)$$

which is exactly the desired conclusion for the point y_0. □

Example 8.17. We apply Theorem 8.15 to our previous examples. We write the ω explicitly to make clear which quantities are random variables and which are not.

(i) For $(X_1, X_2, X_3) \sim \text{Mult}(n, 3, p_1, p_2, p_3)$, Example 8.2 derived

$$E[X_2 \,|\, X_1 = m] = (n - m)\frac{p_2}{p_2 + p_3}.$$

Hence the function v is $v(m) = (n - m)\frac{p_2}{p_2+p_3}$, and consequently

$$E[X_2 \,|\, X_1](\omega) = v(X_1(\omega)) = (n - X_1(\omega))\frac{p_2}{p_2 + p_3}.$$

(ii) For the sum $Z = X + Y$ of two independent $\text{Exp}(\lambda)$ random variables, Example 8.5 found the conditional expectation of X, given $Z = z$, to be $E[X|Z = z] = \frac{1}{2}z$. In this case $v(z) = \frac{1}{2}z$ and

$$E[X|Z](\omega) = v(Z(\omega)) = \tfrac{1}{2}Z(\omega).$$

(iii) For the sum $Z = X + Y$ of two independent Poisson variables $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, Example 8.10 found the conditional expectation of X, given $Z = \ell$, to be $E[X|Z = \ell] = \frac{\lambda}{\lambda+\mu}\ell$. So now $v(\ell) = \frac{\lambda}{\lambda+\mu}\ell$ and

$$E[X|Z](\omega) = v(Z(\omega)) = \tfrac{\lambda}{\lambda+\mu}Z(\omega).$$

△

Taking the function h to be the constant 1 in (8.18) gives the identity

$$E[\,E(X \,|\, Y)\,] = E[X]. \tag{8.22}$$

In other words, the expectation of the conditional expectation is the unconditioned expectation. The next two examples illustrate how this identity can simplify computation of expectations.

Example 8.18. You hold a stick of unit length. Someone comes along and breaks off a uniformly distributed random piece. Now you hold a stick of length Y. Another person comes along and breaks off another uniformly distributed piece from the remaining part of the stick that you hold. You are left with a stick of length X. Find the probability density function f_X, mean $E(X)$ and variance $\text{Var}(X)$ of X.

The problem tells us that $Y \sim \text{Unif}(0, 1)$, and conditional on $Y = y$, $X \sim \text{Unif}(0, y)$. Thus,

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & y \le 0 \text{ or } y \ge 1, \end{cases} \qquad \text{and} \qquad f_{X|Y}(x|y) = \begin{cases} \frac{1}{y}, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The density function of X is, for $0 < x < 1$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)\, f_Y(y)\, dy = \int_x^1 \frac{1}{y}\, dy = -\ln x.$$

Note the integration limits that come from the case-by-case formula for $f_{X|Y}(x|y)$. For $x \notin (0, 1)$ we have $f_X(x) = 0$.

We could compute the expectation and variance of X with the just computed density function f_X, but it is easier to use conditional expectations. Let $0 < y < 1$.

Since the mean of a uniform random variable on an interval is the midpoint of the interval, $E(X|Y=y) = \frac{1}{2}y$. Next compute

$$E(X^2|Y=y) = \int_{-\infty}^{\infty} x^2 \, f_{X|Y}(x|y)\, dx = \int_0^y \frac{x^2}{y}\, dx = \tfrac{1}{3}y^2.$$

From these we deduce first $E(X|Y) = \frac{1}{2}Y$ and $E(X^2|Y) = \frac{1}{3}Y^2$. Then, by taking expectations of the conditional expectations,

$$E(X) = E[E(X|Y)] = \tfrac{1}{2}E(Y) = \tfrac{1}{4},$$

since $E[Y] = \frac{1}{2}$, and

$$E(X^2) = E[E(X^2|Y)] = \tfrac{1}{3}E(Y^2) = \tfrac{1}{3}\int_0^1 y^2\, dy = \tfrac{1}{9}.$$

Finally, $\mathrm{Var}(X) = E(X^2) - (E[X])^2 = \frac{1}{9} - \frac{1}{16} = \frac{7}{144} \approx 0.049$. Exercise 8.7 asks you to break off a third piece of the stick. △

Example 8.19. We roll a fair die repeatedly. How many fives do we see on average before seeing the first six? To be specific, let N denote the number of rolls needed to see the first six and Y the number of fives in the first $N-1$ rolls. We want $E[Y]$. We calculate this expectation first by conditioning on N, and then again by a slick symmetry argument.

N is the familiar geometric random variable: $P(N=n) = (\frac{5}{6})^{n-1}\frac{1}{6}$ for $n \geq 1$ and $E(N) = 6$. The joint distribution of Y and N is readily deduced by thinking about the individual rolls:

$$\begin{aligned} P(Y=m, N=n) &= P(m \text{ fives and no sixes in the first } n-1 \text{ rolls, six at roll } n) \\ &= \binom{n-1}{m}\left(\tfrac{1}{6}\right)^m \left(\tfrac{4}{6}\right)^{n-1-m} \cdot \tfrac{1}{6}. \end{aligned}$$

The derivation above makes sense for $0 \leq m < n$. The conditional probability mass function of Y given $N=n$ is therefore

$$\begin{aligned} p_{Y|N}(m\,|\,n) &= \frac{P(Y=m, N=n)}{P(N=n)} = \frac{\binom{n-1}{m}\left(\frac{1}{6}\right)^m\left(\frac{4}{6}\right)^{n-1-m}\cdot\frac{1}{6}}{\left(\frac{5}{6}\right)^{n-1}\frac{1}{6}} \\ &= \binom{n-1}{m}\left(\tfrac{1}{5}\right)^m\left(\tfrac{4}{5}\right)^{n-1-m}, \qquad 0 \leq m \leq n-1. \end{aligned} \tag{8.23}$$

Thus given $N=n$, the conditional distribution of Y is $\mathrm{Bin}(n-1, \frac{1}{5})$. If you did not foresee this before the computation, does it seem obvious afterwards?

From knowing the conditional distribution and the mean of a binomial,

$$E[Y\,|\,N=n] = \tfrac{1}{5}(n-1).$$

Hence $E(Y\,|\,N) = \frac{1}{5}(N-1)$ and using (8.18) we get

$$E[Y] = E[E[Y\,|\,N]] = E[\tfrac{1}{5}(N-1)] = \tfrac{1}{5}(E[N]-1) = \tfrac{1}{5}(6-1) = 1.$$

This problem can be solved quickly by utilizing symmetry arguments, without conditioning. For $1 \leq i \leq 5$ let Y_i denote the number of i's in the first $N-1$ rolls. Then

$$N = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 + 1$$

from which, using symmetry in the labels,

$$6 = E[N] = E[Y_1] + E[Y_2] + E[Y_3] + E[Y_4] + E[Y_5] + 1 = 5E[Y_i] + 1$$

and we conclude that $E[Y_i] = 1$ for each $1 \le i \le 5$. △

Linearity of conditional expectation.

Theorem 8.20. *Let α, β be real numbers and X and Y random variables on $(\Omega, \mathcal{F}, P)$ with finite expectation. Let Z be another random variable on $(\Omega, \mathcal{F}, P)$. Then*

$$E[\alpha X + \beta Y \mid Z] = \alpha E[X \mid Z] + \beta E[Y \mid Z] \tag{8.24}$$

in the sense that the random variables on either side of the equality sign are equal with probability one.

Proof. The proof requires us to show that the right-hand side of (8.24) fulfills the definition of the conditional expectation $E[\alpha X + \beta Y \mid Z]$, as specified in Definition 8.14. By their definitions, $E[X \mid Z]$ and $E[Y \mid Z]$ are functions of Z, hence so is the linear combination $\alpha E[X \mid Z] + \beta E[Y \mid Z]$. Then we need to check that this linear combination satisfies (the analogue of) identity (8.17) for all bounded Borel functions h. This will come from the linearity of expectation and the analogous properties of $E[X \mid Z]$ and $E[Y \mid Z]$.

$$\begin{aligned} E\big[\,\big(\alpha E[X \mid Z] + \beta E[Y \mid Z]\,\big)\, h(Z)\,\big] &= \alpha E\big[E[X \mid Z]\, h(Z)\big] + \beta E\big[E[Y \mid Z]\, h(Z)\big] \\ &= \alpha E[X h(Z)] + \beta E[Y h(Z)] = E[\,(\alpha X + \beta Y)\, h(Z)]. \end{aligned}$$

□

Conditioning and independence. Recall from Section 3.1 that two discrete random variables X and Y are independent if and only if

$$p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$$

for all values x, y of the random variables, and two jointly continuous random variables X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$$

for all real x, y. On the other hand, we now have the identities

$$p_{X,Y}(x,y) = p_{X|Y}(x \mid y)\, p_Y(y)$$

and

$$f_{X,Y}(x,y) = f_{X|Y}(x \mid y)\, f_Y(y)$$

that are true without any assumptions.

Comparisons of the right-hand sides of these equations give the following criteria for independence in terms of conditional distributions. These statements amount to saying that X and Y are independent if and only if conditioning on Y does not alter the distribution of X.

Theorem 8.21. *Discrete random variables X and Y are independent if and only if $p_{X|Y}(x \mid y) = p_X(x)$ for all possible values x of X, whenever $p_Y(y) > 0$.*

Jointly continuous random variables X and Y are independent if and only if $f_{X|Y}(x \mid y) = f_X(x)$ for all x, whenever $f_Y(y) > 0$.

Apply this to the computation of conditional expectations. The conclusion is that, for independent X and Y, and for any function g and all y for which the expectations below make sense,

$$E[g(X) \,|\, Y = y] = E[g(X)] \qquad \text{and} \qquad E[g(X)|Y] = E[g(X)]. \tag{8.25}$$

In particular, when X and Y are independent, $E[g(X)|Y]$ is no longer random, but a constant equal to $E[g(X)]$.

We can also deduce this from the abstract definition. Define v as the constant function $v(y) = E[g(X)]$. Then for any bounded Borel function h,

$$E[v(Y)h(Y)] = E\big[E[g(X)]h(Y)\big] = E[g(X)]E[h(Y)] = E[g(X)h(Y)].$$

The justification of the steps above goes as follows. The first step is the definition of v. The second step moves the constant $E[g(X)]$ outside the expectation. The third step uses the independence of X and Y. The equality above amounts to verifying (8.17), and we can conclude that $E[g(X)|Y] = E[g(X)]$.

Conditioning on the random variable itself. Conditioning X on Y that is independent of X is an extreme situation where the conditioning information makes no difference. The opposite extreme would be to condition X on X itself. The outcome is the following: for all random variables X,

$$E[g(X)|X] = g(X). \tag{8.26}$$

This follows immediately from Definition 8.14. Define the function v as g. Then the identity

$$E[v(X)h(X)] = E[g(X)h(X)]$$

is immediate, and verifies (8.26).

We can also use probability mass functions to verify (8.26) in the discrete case. Apply definition (8.1) to the case $Y = X$ to get, when $p_X(y) > 0$,

$$\begin{aligned} p_{X|X}(x \,|\, y) &= P(X = x|X = y) = \frac{P(X = x, X = y)}{P(X = y)} \\ &= \begin{cases} 0, & x \neq y \\ \dfrac{P(X = y, X = y)}{P(X = y)} = \dfrac{P(X = y)}{P(X = y)} = 1, & x = y. \end{cases} \end{aligned}$$

Then from (8.8)

$$E[g(X)|X = y] = \sum_x g(x)\, p_{X|X}(x \,|\, y) = g(y) \cdot 1 = g(y),$$

where the second equality holds because the only nonzero term in the sum over x is the one with $x = y$. Thus $E[g(X)|X = y] = g(y)$, and so by definition, $E[g(X)|X] = g(X)$. In particular, taking the identity function $g(x) = x$ gives the identity

$$E(X|X) = X. \tag{8.27}$$

Example 8.22. Suppose $X_1, \dots, X_n$ are independent and identically distributed and $S_n = X_1 + \dots + X_n$. Find $E(X_i|S_n)$ for $1 \le i \le n$.

Exchangeability implies that $E(X_i|S_n) = E(X_1|S_n)$ for each $1 \le i \le n$. We can see this through the definition:

$$E\big[E(X_1|S_n)\,h(S_n)\big] = E[X_1 h(S_n)] = E[X_i h(S_n)].$$

In the last step we applied a permutation that exchanges i with 1. The equalities show that $E(X_1|S_n)$ satisfies the definition of $E(X_i|S_n)$.

Then by (8.27) and the linearity of conditional expectation,

$$\begin{aligned} S_n &= E(S_n|S_n) = E[X_1 + \dots + X_n|S_n] \\ &= E(X_1|S_n) + \dots + E(X_n|S_n) = nE(X_1|S_n). \end{aligned}$$

Hence $E(X_i|S_n) = E(X_1|S_n) = \frac{S_n}{n}$. △

Conditioning on a random variable fixes its value. We take identity (8.26) a step further by conditioning a function of multiple random variables. Conditioning on $Y = y$ fixes the value of Y to be y. Then Y is no longer random and any occurrence of Y inside the conditional expectation can be replaced with the particular value y. In mathematical terms, for any function $h(x, y)$ we have

$$E[h(X,Y)\,|\,Y=y] = E[h(X,y)\,|\,Y=y]. \tag{8.28}$$

The formula above is valid in general. It is justified as follows in the discrete case. Assume $P(Y = y) > 0$.

$$\begin{aligned} E[h(X,Y)\,|\,Y=y] &= \sum_{x,w} h(x,w) P(X=x, Y=w\,|\,Y=y) \\ &= \sum_{x,w} h(x,w) \frac{P(X=x,\, Y=w,\, Y=y)}{P(Y=y)} \\ &= \sum_{x} h(x,y) \frac{P(X=x,\, Y=y)}{P(Y=y)} \\ &= \sum_{x} h(x,y) P(X=x\,|\,Y=y) = E[h(X,y)\,|\,Y=y]. \end{aligned}$$

In the third equality above, we used that $P(X = x,\, Y = w,\, Y = y) = 0$ unless $w = y$.

Here is an interesting special case. Take two single-variable functions a and b and apply (8.28) to $h(x, y) = a(x)b(y)$. Then

$$E[a(X)b(Y)\,|\,Y=y] = E[a(X)b(y)\,|\,Y=y] = b(y)\,E[a(X)\,|\,Y=y].$$

In the second step we took the constant $b(y)$ out of the expectation. If we switch to conditional expectations as random variables, the formula above becomes

$$E[\,a(X)\,b(Y)\,|\,Y] = b(Y)\,E[\,a(X)\,|\,Y]. \tag{8.29}$$

This equation can also be proved conveniently with Definition 8.14. Namely define the function v by $v(y) = b(y)E[\,a(X)\,|\,Y = y]$. Then for any bounded Borel function h,

$$\begin{aligned} E[v(Y)h(Y)] &= E\big[b(Y)\,E[\,a(X)\,|\,Y]\,h(Y)\big] = E\big[\,E[\,a(X)\,|\,Y]\,b(Y)\,h(Y)\big] \\ &= E[a(X)\,b(Y)\,h(Y)] \end{aligned}$$

which verifies that $v(Y) = E[\,a(X)\,b(Y)\,|\,Y]$. The last step above is identity (8.17) aplied to the function $\widetilde{h}(y) = b(y)h(y)$.

It seems like we broke the rules in (8.29) by moving the random variable $b(Y)$ outside the expectation. But this is not an ordinary expectation, it is a conditional expectation that is itself a random variable. *By conditioning on Y, we are allowed to treat $b(Y)$ as if it were a constant.*

Example 8.23. Recall the multinomial setting from Example 8.2. A trial has outcomes $1, 2$, and 3 that occur with probabilities p_1, p_2, and p_3, respectively. The random variable X_i is the number of times outcome i appears among n independent repetitions of the trial. We give a new computation of $\text{Cov}(X_1, X_2)$, previously done in Example 4.32.

First recall that the marginal distribution of each X_i is $\text{Bin}(n, p_i)$. Hence,

$$E[X_i] = np_i \quad \text{and} \quad E[X_i^2] = \text{Var}(X_i) + (E[X_i])^2 = np_i(1-p_i) + n^2p_i^2.$$

In Example 8.2 we deduced $E[X_2\,|\,X_1 = m] = (n-m)\frac{p_2}{p_2+p_3}$, from which we get

$$E[X_2\,|\,X_1] = (n - X_1)\frac{p_2}{p_2+p_3}.$$

We may now compute as follows,

$$\begin{aligned} E[X_1X_2] &= E[\,E(X_1X_2|X_1)\,] = E[\,X_1\,E(X_2|X_1)\,] = E\big[X_1(n-X_1)\tfrac{p_2}{p_2+p_3}\big] \\ &= \tfrac{p_2}{p_2+p_3}\big(nE[X_1] - E[X_1^2]\big) = -np_1p_2 + n^2p_1p_2. \end{aligned}$$

The first equality above follows from (8.22), the second from an application of (8.29), and the third from the above expression for $E[X_2\,|\,X_1]$. We also used that $p_1 + p_2 + p_3 = 1$, but omitted the algebra steps. From this we get the covariance:

$$\begin{aligned} \text{Cov}(X_1, X_2) &= E[X_1X_2] - E[X_1]E[X_2] = -np_1p_2 + n^2p_1p_2 - np_1np_2 \\ &= -np_1p_2. \end{aligned}$$

△

8.3. Additional conditioning examples

This section develops further topics on conditioning and illustrates computational techniques. Individual parts of this section can be read independently of each other. We compute a moment generating function by conditioning, illustrate mixing discrete and continuous random variables, show how the conditional expectation provides the best estimate in a least-squares sense, and study random sums. After this, the last two parts each take a small step towards broad subjects that are natural follow-ups to an introductory probability course: statistics and stochastic processes.

Conditional moment generating function. The example below identifies an unknown probability mass function by finding its moment generating function. The moment generating function is computed by conditioning.

Example 8.24. We continue with the setting of Example 8.19. N is the number of rolls of a fair die needed to see the first six and Y is the number of fives in the first $N-1$ rolls. The task is to find the probability mass function $p_Y(m) = P(Y = m)$. We compute the moment generation function $M_Y(t)$ of Y by conditioning on N. Let $t \in \mathbb{R}$.

$$M_Y(t) = E[e^{tY}] = E\big[E[e^{tY} \,|\, N]\big] = \sum_{n=1}^{\infty} E[e^{tY} \,|\, N = n]\, P(N = n). \tag{8.30}$$

To compute the conditional expectation of e^{tY} needed above, we take the conditional probability mass function of Y from (8.23). The binomial theorem evaluates the sum below in a tidy formula.

$$\begin{aligned} E[e^{tY} \,|\, N = n] &= \sum_{m=0}^{n-1} e^{tm}\, p_{Y|N}(m \,|\, n) = \sum_{m=0}^{n-1} e^{tm} \binom{n-1}{m} \big(\tfrac{1}{5}\big)^m \big(\tfrac{4}{5}\big)^{n-1-m} \\ &= \big(\tfrac{1}{5}e^t + \tfrac{4}{5}\big)^{n-1}. \end{aligned}$$

Substitute this and $P(N = n) = \big(\frac{5}{6}\big)^{n-1}\frac{1}{6}$, for $n \ge 1$, into (8.30), rearrange, and apply the formula for a geometric series to find

$$\begin{aligned} E[e^{tY}] &= \sum_{n=1}^{\infty} \big(\tfrac{1}{5}e^t + \tfrac{4}{5}\big)^{n-1} \cdot \big(\tfrac{5}{6}\big)^{n-1}\tfrac{1}{6} = \frac{1}{6}\sum_{n=1}^{\infty}\left(\frac{4+e^t}{6}\right)^{n-1} \\ &= \tfrac{1}{6} \cdot \frac{1}{1 - \frac{4+e^t}{6}} = \frac{1}{2 - e^t}. \end{aligned} \tag{8.31}$$

In order for the series to converge we must have $\frac{4+e^t}{6} < 1$, which is equivalent to $t < \ln 2$.

We must now find the distribution whose moment generating function is $\frac{1}{2-e^t}$. Since the values of Y are nonnegative integers, its moment generating function is of the form $M_Y(t) = \sum_{k=0}^{\infty} P(Y = k)e^{tk}$. Expanding the function from (8.31) with the geometric series formula $(1-x)^{-1} = \sum_{k=0}^{\infty} x^k$ yields

$$\frac{1}{2 - e^t} = \frac{1}{2} \cdot \frac{1}{1 - e^t/2} = \frac{1}{2}\sum_{k=0}^{\infty}\left(\frac{e^t}{2}\right)^k = \sum_{k=0}^{\infty}\big(\tfrac{1}{2}\big)^{k+1} e^{tk}. \tag{8.32}$$

Since the moment generating function uniquely determines the distribution, we conclude that $P(Y = k) = (\frac{1}{2})^{k+1}$ for $k \in \{0, 1, 2, \dots\}$. Thus $Y = X - 1$ for $X \sim \text{Geom}(1/2)$. Y is called a *shifted geometric variable.*

What is the intuitive explanation for $P(Y = k) = \big(\frac{1}{2}\big)^{k+1}$? If we discard rolls of $1, 2, 3,$ and 4 and record only fives and sixes, then fives and sixes come with probability $\frac{1}{2}$ each. Among these recorded rolls the position of the first six is a Geom(1/2) random variable. All the earlier recorded rolls are fives. Thus the number of fives is a Geom(1/2) random variable minus one. △

Mixing discrete and continuous random variables. The conditioning machinery of this chapter is more flexible than is initially obvious. In particular, we can comfortably mix discrete and continuous random variables. This example illustrates.

Example 8.25 (Conditioning a continuous random variable on a discrete random variable)**.** A factory produces lightbulbs with two different machines. Bulbs from machine 1 have an exponential lifetime with parameter $\lambda > 0$, while bulbs from machine 2 have an exponential lifetime with parameter $\mu > 0$. Suppose machine 1 produces 2/3 of the lightbulbs. Let T be the lifetime of a randomly chosen bulb from this factory. Find the probability density function f_T of T.

Our intuition tells us that the answer must be $f_T(t) = \frac{2}{3}\lambda e^{-\lambda t} + \frac{1}{3}\mu e^{-\mu t}$, for $t \geq 0$, and zero for $t < 0$. Let us see how this can be shown.

First note that for $t < 0$, we have $F_T(t) = P(T \leq t) = 0$. Hence $f_T(t) = \frac{d}{dt}F_T(t) = 0$.

Now we consider $t \geq 0$. Let $Y \in \{1, 2\}$ denote the machine that produced the bulb. The problem statement gives us the following information:

$$P(Y = 1) = \tfrac{2}{3}\,, \quad P(Y = 2) = \tfrac{1}{3}\,, \quad f_{T|Y}(t\,|\,1) = \lambda e^{-\lambda t} \quad \text{and} \quad f_{T|Y}(t\,|\,2) = \mu e^{-\mu t},$$

for $t \geq 0$. We can find the density function f_T by first finding the cumulative distribution function and then differentiating. We do this by conditioning on Y. For $t \geq 0$ we have

$$\begin{aligned} F_T(t) &= P(T \leq t) = P(Y = 1)P(T \leq t\,|\,Y = 1) + P(Y = 2)P(T \leq t\,|\,Y = 2) \\ &= P(Y = 1)\int_{-\infty}^{t} f_{T|Y}(s|1)\,ds + P(Y = 2)\int_{-\infty}^{t} f_{T|Y}(s|2)\,ds \\ &= \tfrac{2}{3}\int_0^t \lambda e^{-\lambda s}\,ds + \tfrac{1}{3}\int_0^t \mu e^{-\mu s}\,ds \\ &= \int_0^t \left(\tfrac{2}{3}\lambda e^{-\lambda s} + \tfrac{1}{3}\mu e^{-\mu s}\right) ds. \end{aligned}$$

We do not need to evaluate the integral since by the fundamental theorem of calculus,

$$f_T(t) = \frac{d}{dt}F_T(t) = \tfrac{2}{3}e^{-\lambda t} + \tfrac{1}{3}\mu e^{-\mu t} \qquad \text{for } t > 0. \tag{8.33}$$

What is the mean lifetime $E[T]$? Since the mean of an $\text{Exp}(\lambda)$ random variable is λ^{-1}, we have $E[T|Y = 1] = \lambda^{-1}$ and $E[T|Y = 2] = \mu^{-1}$. Consequently

$$E[T] = E[E[T|Y]] = \tfrac{2}{3}E[T|Y = 1] + \tfrac{1}{3}E[T|Y = 2] = \tfrac{2}{3}\lambda^{-1} + \tfrac{1}{3}\mu^{-1}.$$

Of course, this could also have been computed directly from our probability density function (8.33).

△

Conditional expectation as the best predictor. Suppose we wish to know a random quantity X, but we cannot observe X directly. What we know is another random variable Y, perhaps an inaccurate measurement corrupted by random noise. What is the best estimate of X in terms of Y? An estimate of X based on the value of Y must be $h(Y)$ for some function h. We measure the effectiveness of an estimate with the expected square error $E[(X-h(Y))^2]$. The theorem below states that in this sense the uniquely best estimate is the conditional expectation $h(Y)=E(X\,|\,Y)$.

Theorem 8.26. *Assuming that the expectations below are well-defined, the inequality*

$$E\big[\big(X-h(Y)\big)^2\big]\geq E\big[\big(X-E[X\,|\,Y]\big)^2\big]$$

holds for any function h. Equality holds above only for the choice $h(Y)=E[X|Y]$.

Proof. The theorem is true completely generally. We prove it assuming that X,Y are jointly continuous. Exercise 8.8 asks for the proof for the discrete case.

For an arbitrary function h, add and subtract $E[X\,|\,Y]$ inside $E[(X-h(Y))^2]$ and expand the square:

$$\begin{aligned}E\big[\big(X-h(Y)\big)^2\big]&=E\big[\big(X-E[X\,|\,Y]+E[X\,|\,Y]-h(Y)\big)^2\big]\\&=E\big[\big(X-E[X\,|\,Y]\big)^2\big]+E\big[\big(E[X\,|\,Y]-h(Y)\big)^2\big]\\&\quad+2E\big[\big(X-E[X\,|\,Y]\big)\big(E[X\,|\,Y]-h(Y)\big)\big].\end{aligned}\tag{8.34}$$

We check that the expectation on line (8.34) vanishes. It equals

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big(x-E[X\,|\,Y=y]\big)\big(E[X\,|\,Y=y]-h(y)\big)\,f_{X,Y}(x,y)\,dx\,dy\\&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big(x-E[X\,|\,Y=y]\big)\big(E[X\,|\,Y=y]-h(y)\big)\,f_{X|Y}(x\,|\,y)\,f_Y(y)\,dx\,dy\\&=\int_{-\infty}^{\infty}\Big[\int_{-\infty}^{\infty}\big(x-E[X\,|\,Y=y]\big)f_{X|Y}(x\,|\,y)\,dx\Big]\big(E[X\,|\,Y=y]-h(y)\big)\,f_Y(y)\,dy.\end{aligned}$$

The inner x-integral is calculated under a fixed y. It equals

$$\int_{-\infty}^{\infty}x\,f_{X|Y}(x\,|\,y)\,dx-E[X\,|\,Y=y]\;=\;E[X\,|\,Y=y]-E[X\,|\,Y=y]\;=\;0.$$

Consequently the term on line (8.34) is zero, and we are left with the equality

$$E\big[\big(X-h(Y)\big)^2\big]=E\big[\big(X-E[X\,|\,Y]\big)^2\big]+E\big[\big(E[X\,|\,Y]-h(Y)\big)^2\big].\tag{8.35}$$

Because the last term $E\big[\big(E[X\,|\,Y]-h(Y)\big)^2\big]$ above is nonnegative, equation (8.35) tells us that the mean square error from any approximation $h(Y)$ is at least as large as the mean square error when X is approximated by $E[X\,|\,Y]$. That is,

$$E\big[\big(X-h(Y)\big)^2\big]\geq E\big[\big(X-E[X\,|\,Y]\big)^2\big],$$

which is what we had set out to show.

To see when the inequality is strict, observe that the last expectation on line (8.35) equals

$$\int_{-\infty}^{\infty}\big(E(X\,|\,Y=y)-h(y)\big)^2f_Y(y)\,dy.$$

Unless $h(y) = E(X \mid Y = y)$ for all y such that $f_Y(y) > 0$, this integral is strictly positive and leads to

$$E\big[\big(X - h(Y)\big)^2\big] \;>\; E\big[\big(X - E(X \mid Y)\big)^2\big]. \qquad \square$$

Random sums. Suppose the average size of a claim on a certain car insurance policy is \$1,000. Suppose further that on average the insurance company pays 20 claims per year. What is the average total amount paid out during a year? Linearity of expectation would seem to give the answer $20 \times \$1,000 = \$20,000$. However, if *both* the sizes of individual claims *and* the number of claims are *random*, simple linearity does not apply. We need a new theorem called Wald's identity.

In the general setting $X_1, X_2, X_3, \dots$ are i.i.d. random variables with finite mean and $S_n = X_1 + \cdots + X_n$, with $S_0 = 0$. Let N be a nonnegative integer valued random variable with a finite expectation that is independent of the random variables X_i. Let $S_N = X_1 + \cdots + X_N$. This means that N tells us how many terms to add up from the sequence of X_is. The convention is that if $N = 0$ then $S_N = 0$. S_N is a *random sum* because both the individual terms and the number of terms are random. To connect with the question posed in the paragraph above, the random variables X_i are the values of the claims and N is the number of claims in a given year. Here is the formula for the mean of S_N.

Theorem 8.27 (Wald's identity)**.** *Let $X_1, X_2, X_3 \dots$ be i.i.d. random variables with finite mean. Let N be a nonnegative integer-valued random variable independent of the X_is, also with finite mean. Let $S_N = X_1 + \cdots + X_N$. Then*

$$E[S_N] = E[N] \cdot E[X_1].$$

The qualitative conclusion of Wald's identity is that linearity of expectation works in an averaged sense: the expected value of the random sum equals the expected value of the number of terms times the expected value of a single term. It is important to note that

$$E\Big[\sum_{k=1}^{N} X_k\Big] \neq \sum_{k=1}^{N} E[X_k] \qquad \text{when } N \text{ is random.}$$

The left-hand side is a number (an expectation) while the right-hand side is a random variable if N is random.

We give a proof of Wald's identity with the help of conditional expectations.

Proof of Wald's identity. Let $\mu = E[X_i]$. Conditioning on $\{N = n\}$ yields

$$E[S_N|N = n] = E[S_n|N = n] = E[S_n] = E[X_1] + \cdots + E[X_n] = n\mu.$$

First we replaced S_N with S_n since we conditioned on $N = n$. Next we dropped the conditioning due to the independence of N and S_n. Finally, linearity of expectation applies because the number of terms is a constant n.

The equality above turns into $E[S_N|N] = N\mu$, and then we compute the expectation: $E[S_N] = E[\,E[S_N|N]\,] = E[N\mu] = \mu E[N]$. $\square$

Example 8.28. Suppose the number of small accidents per year experienced by a fleet of cars is Poisson distributed with mean 20. 90% of the accidents require only a paint job that costs \$100, while 10% of the accidents need bodywork that costs \$5000. What is the average annual cost of these accidents? Assume that the costs of individual accidents and the number of accidents are all independent random variables.

The total cost is the random sum $S_N = X_1 + \cdots + X_N$ where $N \sim \text{Poisson}(20)$ and each X_i has probability mass function $P(X_i = 100) = 0.9$ and $P(X_i = 5000) = 0.1$. By Wald's identity,

$$E[S_N] = E[N] \cdot E[X_1] = 20 \cdot (0.9 \cdot 100 + 0.1 \cdot 5000) = 1180.$$

△

You might wonder whether Wald's identity is true even if N is not independent of the random variables X_i. The answer is "not necessarily", as the following example shows.

Example 8.29 (Breaking Wald's identity)**.** Let $X_1, X_2, X_3, \dots$ be independent Bernoulli random variables with parameter $0 < p < 1$. Let N be the number of failures before the first success. Then $N + 1$ is the position of the first success, and so $N + 1 \sim \text{Geom}(p)$ and $E[N] = \frac{1}{p} - 1 = \frac{1-p}{p}$. We have

$$S_N = X_1 + \cdots + X_N = 0,$$

since $X_i = 0$ for all $i \le N$. This yields $E[S_N] = 0$, which is not equal to

$$E[N] \cdot E[X_1] = \frac{1-p}{p} \cdot p = 1 - p.$$

△

The story of Wald's identity does not end here. Independence of N from $\{X_i\}$ is in fact not necessary for Wald's identity, as long as N belongs to a class of random variables called *stopping times*. This notion is studied in the theory stochastic processes.

Trials with unknown success probability. In Section 6.5 we used the observed frequency of successes $\widehat{p} = S_n/n$ to estimate the unknown success probability p of independent trials. Use of $\widehat{p}$ was justified by appeal to the law of large numbers and by the observation that $\widehat{p}$ is the maximum likelihood estimator of p (Remark 6.25).

In this section we take a different view of the same problem. We use probability theory to model our ignorance of the success probability and the accumulation of information from observed outcomes. Let $X_i \in \{0, 1\}$ be the outcome of the ith trial and $S_n = X_1 + \cdots + X_n$ the number of successes in n trials. Let ξ denote the unknown success probability. To represent absence of knowledge about ξ, assume that ξ is a uniform random variable on $(0, 1)$. Conditionally on $\xi = p$ the trials are independent with success probability p. Then, given $\xi = p$, we have $P(X_i = 1 \,|\, \xi = p) = p = 1 - P(X_i = 0 \,|\, \xi = p)$ and $S_n \sim \text{Bin}(n, p)$ with conditional probability mass function

$$P(S_n = k \,|\, \xi = p) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for } k = 0, 1, \dots, n. \tag{8.36}$$

The next three examples answer questions about this model.

Example 8.30 (Unconditional probability distribution of successes)**.** First we observe that the unconditional success probability of each trial is $\frac{1}{2}$. Let f_ξ denote the density function of ξ, given by $f_\xi(p) = 1$ for $0 < p < 1$.

$$P(X_i = 1) = \int_0^1 P(X_i = 1 \,|\, \xi = p) f_\xi(p)\, dp = \int_0^1 p\, dp = \tfrac{1}{2}. \tag{8.37}$$

To determine the distribution of S_n we compute its moment generating function $M_{S_n}(t)$ by conditioning on ξ. Let $t \neq 0$ (we already know that $M_{S_n}(0) = 1$).

$$\begin{aligned} M_{S_n}(t) = E[e^{tS_n}] &= \int_0^1 E[e^{tS_n} \,|\, \xi = p] f_\xi(p)\, dp = \int_0^1 \sum_{k=0}^n \binom{n}{k} e^{tk} p^k (1-p)^{n-k}\, dp \\ &= \int_0^1 (pe^t + 1 - p)^n\, dp = \frac{1}{e^t - 1} \int_1^{e^t} u^n\, du = \frac{1}{e^t - 1} \cdot \frac{u^{n+1}}{n+1}\bigg|_{u=1}^{u=e^t} \\ &= \frac{1}{n+1} \cdot \frac{e^{(n+1)t} - 1}{e^t - 1}. \end{aligned}$$

We used the binomial theorem and then the substitution $u = pe^t + 1 - p$ with $du = (e^t - 1)dp$.

Next we find the probability mass function whose moment generating function is the one above. Formula (C.7) for finite geometric sums gives us

$$\frac{1}{n+1} \cdot \frac{e^{(n+1)t} - 1}{e^t - 1} = \frac{1}{n+1} \sum_{k=0}^n e^{tk}.$$

Since a moment generating function (finite in an interval around 0) determines the probability distribution uniquely, we conclude that $P(S_n = k) = \frac{1}{n+1}$ for $k = 0, 1, \dots, n$. In other words, S_n is a uniform random variable on $\{0, 1, \dots, n\}$. (Exercise 8.10 gives an alternative route to this answer.)

The distribution of S_n is dramatically different from a binomial. As a consequence, the trials are *not* independent without the conditioning on ξ. If they were unconditionally independent, (8.37) would imply that $S_n \sim \text{Bin}(n, \frac{1}{2})$. However, the trials are exchangeable (Exercise 8.9). △

Example 8.31 (Estimating the unknown success probability)**.** Repetitions of the trial give us information about the success probability. Given that we have seen k successes in n trials, what is the conditional density function of ξ?

While we do not have a definition of a conditional density function $f_{\xi|S_n}(p|k)$ that mixes discrete and continuous, we can proceed by analogy with our previous theory. Guided by equation (8.5), the conditional density function should give conditional probabilities, and so we expect the equality

$$P(a \le \xi \le b \,|\, S_n = k) = \int_a^b f_{\xi|S_n}(p|k)\, dp.$$

We can compute the left-hand side by moving the conditioning from S_n to ξ:

$$\begin{aligned} P(a \le \xi \le b \,|\, S_n = k) &= \frac{P(a \le \xi \le b,\, S_n = k)}{P(S_n = k)} \\ &= \frac{1}{P(S_n = k)} \int_a^b P(S_n = k \,|\, \xi = p)\, f_\xi(p)\, dp \\ &= (n+1) \int_a^b \binom{n}{k} p^k (1-p)^{n-k}\, dp. \end{aligned}$$

We conclude that the conditional density function of ξ given $S_n = k$ is

$$f_{\xi|S_n}(p|k) = \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k} \quad \text{for} \quad 0 < p < 1. \tag{8.38}$$

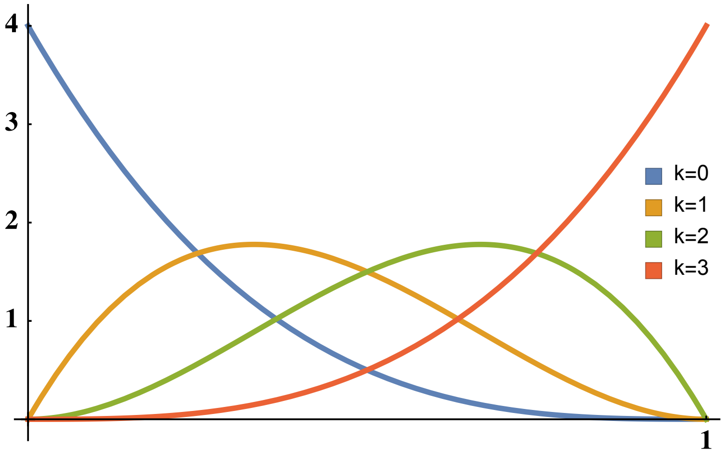

Figure 2. The conditional density function of ξ when $n = 3$ and k is 0, 1, 2 or 3. As the number of successes increases, the peak of the density function moves to the right.

The density function discovered above for ξ given $S_n = k$ is the beta distribution with parameters $(k+1, n-k+1)$. (Figure 2 illustrates.) The general Beta(a, b) density function with parameters $a, b > 0$ is defined as

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \qquad \text{for } 0 < x < 1 \tag{8.39}$$

and zero otherwise. Recall that $\Gamma(t)$ is the gamma function defined for real $t > 0$ as

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx,$$

and $\Gamma(n) = (n-1)!$ for positive integers n.

△

Remark 8.32 (Bayesian versus classical statistics)**.** Remark 6.25 and Example 8.31 illustrate a difference between *classical* and *Bayesian* statistics.

In classical statistics the unknown success probability p is regarded as a fixed quantity. Based on the information from the trials we formed the maximum likelihood estimator $\widehat{p} = k/n$ for p, where k is the observed number of successes.

In the Bayesian approach a prior probability distribution is placed on the unknown parameter, and then the posterior distribution is calculated with Bayes' rule. In the example above, the uniform distribution on ξ is the prior distribution, and the Beta$(k+1, n-k+1)$ distribution obtained in (8.38) is the posterior. The conditional expectation of ξ gives a Bayesian point estimate of p:

$$E(\xi|S_n = k) = \int_0^1 p\, f_{\xi|S_n}(p|k)\, dp = \frac{(n+1)!}{k!(n-k)!} \int_0^1 p^{k+1}(1-p)^{n-k}\, dp = \frac{k+1}{n+2}.$$

This calculation takes into account both the observed frequency k/n and the prior distribution, and hence does not agree with the maximum likelihood estimate $\widehat{p}$. △

Example 8.33 (Success probability of the next trial)**.** Suppose we had k successes in the first n trials. What is then $P(X_{n+1} = 1|S_n = k)$, the conditional probability that the next experiment is a success?

By the definition of conditional probability

$$P(X_{n+1} = 1|S_n = k) = \frac{P(X_{n+1} = 1, S_n = k)}{P(S_n = k)}.$$

We already computed $P(S_n = k) = \frac{1}{n+1}$. To compute the numerator we condition on the success probability ξ:

$$\begin{aligned} P(X_{n+1} = 1, S_n = k) &= \int_0^1 P(X_{n+1} = 1, S_n = k|\xi = p) f_\xi(p)\, dp \\ &= \int_0^1 P(X_{n+1} = 1|\xi = p) P(S_n = k|\xi = p) f_\xi(p)\, dp \\ &= \int_0^1 p \binom{n}{k} p^k (1-p)^{n-k} dp, \end{aligned} \tag{8.40}$$

where the second step follows from the conditional independence of the random variables $X_1, X_2, \dots$. To evaluate the last integral we use the beta density function (8.39):

$$\begin{aligned} \int_0^1 p\binom{n}{k} p^k(1-p)^{n-k} dp &= \frac{k+1}{(n+1)(n+2)} \int_0^1 \frac{(n+2)!}{(k+1)!(n-k)!} p^{k+1}(1-p)^{n-k} dp \\ &= \frac{k+1}{(n+1)(n+2)}, \end{aligned}$$

where the last integral equals 1 since it is the integral of the Beta$(k+2, n-k+1)$ density function over the real line. From this

$$P(X_{n+1} = 1|S_n = k) = \frac{\frac{k+1}{(n+1)(n+2)}}{\frac{1}{n+1}} = \frac{k+1}{n+2}.$$

△

This last example is sometimes called the *sunrise problem* or *Laplace's law of succession.* In the 18th century Pierre-Simon Laplace computed the probability that the sun rises the next day, given that it had risen every day for the previous 5000 years (which was believed to be the age of the Earth at that time). Laplace assumed that sunrises happen according to independent Bernoulli random variables with a success probability that was “chosen” at the beginning of time according to a uniform distribution.

Exercises

Exercise 8.1. We roll a die until we get a six and denote the number of rolls by X. Then we take a fair coin and we repeatedly flip it until we get X heads. We denote the number of coin flips needed by Y.

(a) Find the conditional probability mass function of Y given $X = x$.

(b) Find the probability mass function of Y.

(c) Find the conditional probability mass function of X given $Y = y$. What kind of distribution is that?

Exercise 8.2. Let X and Y be discrete random variables. Show that

$$p_X(x) = \sum_y p_{X|Y}(x \,|\, y)\, p_Y(y),$$

$$E[g(X) \,|\, Y = y] = \sum_x g(x)\, p_{X|Y}(x \,|\, y)$$

for y such that $P(Y = y) > 0$, and

$$E[g(X)] = \sum_y E[g(X) \,|\, Y = y]\, p_Y(y).$$

Exercise 8.3. Example 8.11 had random variables M, L and the probability mass functions

$$p_M(m) = \binom{n}{m} p^m (1-p)^{n-m} \quad \text{and} \quad p_{L|M}(\ell|m) = \binom{m}{\ell} r^\ell (1-r)^{m-\ell}$$

for integers $0 \le \ell \le m \le n$, with n fixed. Find the marginal probability mass function p_L of the random variable L. Before computing, try to deduce it with common-sense intuition from the description of the example.

Exercise 8.4. We generate a random variable $Y \sim \text{Exp}(1)$. Then we generate another exponential random variable X whose parameter is the value Y just observed.

(a) Find the joint density function of (X, Y).

(b) Suppose we observe $X = x$. How is Y now distributed? Identify the distribution by name.

Exercise 8.5. Let the joint density function of (X, Y) be

$$f(x, y) = \frac{1}{y} e^{-x/y} e^{-y} \quad \text{for } 0 < x < \infty \text{ and } 0 < y < \infty.$$

(a) Find $f_Y(y)$ and $f_{X|Y}(x|y)$. Compute $E[Y]$.

(b) Find the conditional expectation $E[X \,|\, Y]$.

(c) Use parts (a) and (b) to compute $E[X]$.

Exercise 8.6. Let (X, Y) be discrete random variables and g a function such that $g(X)$ has a finite expectation. Show that $E[\, g(X) \,|\, Y](\omega) = v(Y(\omega))$ for the function $v(y) = E[g(X) \,|\, Y = y]$ defined by (8.8).

Exercise 8.7. The exercise continues the theme of Example 8.18. You hold a stick of unit length. Three pieces are broken off one after another, and each time the piece taken away is uniformly distributed on the length of the piece remaining in your hand. Let Z be the length of the piece that remains after the three pieces have been broken off. Find the density function f_Z, mean $E[Z]$ and variance $\mathrm{Var}(Z)$ of Z.

Exercise 8.8. Assume that X, Y are discrete random variables and h a function. Show that unless $h(y) = E(X \,|\, Y = y)$ for all y such that $p_Y(y) > 0$, we have the strict inequality

$$E\big[\big(X - h(Y)\big)^2\,\big] > E\big[\big(X - E[X \,|\, Y]\big)^2\,\big].$$

Exercise 8.9. In the context of Example 8.30 of trials conditionally independent on the success probability $\xi = p$, show that the outcomes $X_1, \dots, X_n$ are exchangeable. Concretely, show that for any choice $t_1, \dots, t_n$ of zeroes and ones,

$$P(X_1 = t_1, X_2 = t_2, \dots, X_n = t_n) = P(X_1 = t_{k_1}, X_2 = t_{k_2}, \dots, X_n = t_{k_n})$$

for any permutation $(k_1, k_2, \dots, k_n)$ of $(1, 2, \dots, n)$.

Exercise 8.10. In the context of the trials with uniformly distributed success probability in Section 8.3, by averaging the conditional probability mass function (8.36) of S_n we have the formula

$$P(S_n = k) = \int_0^1 P(S_n = k \,|\, \xi = p)\, f_\xi(p)\, dp = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k}\, dp.$$

Show by integration by parts that $P(S_n = k) = P(S_n = k + 1)$ for $0 \le k < n$, and conclude from this that $P(S_n = k) = \frac{1}{n+1}$ for $0 \le k \le n$.

Chapter 9

Further topics

9.1. Coupling and the law of rare events

Theorem 3.31 showed that if $\lambda > 0$ is fixed and $S_n \sim \text{Bin}(n, \lambda/n)$ for $n > \lambda$, then as $n \to \infty$, S_n converges in distribution to a Poisson(λ) random variable. Specifically, $P(S_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}$ for each fixed $k \in \mathbb{Z}_{\geq 0}$. The next theorem gives a quantitative error estimate for this convergence. We also generalize the treatment to independent trials whose success probabilities can vary.

Theorem 9.1. *Let $p_1, \dots, p_n \in [0,1]$ and let $X_1, \dots, X_n$ be independent random variables with marginal distributions $X_i \sim \text{Ber}(p_i)$. Let $S = \sum_{i=1}^n X_i$ and $\lambda = \sum_{i=1}^n p_i$. Then*

$$\sum_{k=0}^{\infty} \left| P(S = k) - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq 2 \sum_{i=1}^{n} p_i^2.$$

From this theorem we can derive a more general Poisson limit for independent trials. The two assumptions in the corollary below require that, as $n \to \infty$, (i) the mean number of successes converges and (ii) the success probability declines to zero uniformly.

Corollary 9.2. *For each index $n \in \mathbb{Z}_{>0}$ let m_n be a positive integer and $\{p_{n,i} : 1 \leq i \leq m_n\}$ numbers in $[0,1]$. Consider a sequence of experiments where the nth experiment consists of m_n independent trials with success probabilities $p_{n,1}, p_{n,2}, \dots, p_{n,m_n}$. Let S_n be the number of successes in the nth experiment and $\lambda_n = E(S_n) = \sum_{i=1}^{m_n} p_{n,i}$ the mean number of successes in the nth experiment. Make two assumptions:*

(i) $\lambda_n \to \lambda$.

(ii) $\displaystyle\lim_{n\to\infty} \max_{1 \leq i \leq m_n} p_{n,i} = 0$.

Then S_n converges in distribution to a Poisson(λ) *random variable.*

Proof. Fix $k \in \mathbb{Z}_{\geq 0}$. Apply below the bound from Theorem 9.1.

$$\begin{aligned}&\Big|P(S=k)-\frac{\lambda^k}{k!}e^{-\lambda}\Big|\\ &\leq \Big|P(S=k)-\frac{\lambda_n^k}{k!}e^{-\lambda_n}\Big|+\Big|\frac{\lambda_n^k}{k!}e^{-\lambda_n}-\frac{\lambda^k}{k!}e^{-\lambda}\Big|\\ &\leq 2\sum_{i=1}^{m_n}p_{n,i}^2+\Big|\frac{\lambda_n^k}{k!}e^{-\lambda_n}-\frac{\lambda^k}{k!}e^{-\lambda}\Big|\\ &\leq 2\lambda_n \max_{1\leq i\leq m_n} p_{n,i}+\Big|\frac{\lambda_n^k}{k!}e^{-\lambda_n}-\frac{\lambda^k}{k!}e^{-\lambda}\Big|.\end{aligned}$$

As $n \to \infty$, the last line above converges to zero by the assumptions. □

Proof of Theorem 9.1. The inequality

$$e^x \geq 1+x \tag{9.1}$$

is used several times below. It is valid for all real x, and can be verified with calculus: the global minimum of $f(x)=e^x-1-x$ is $f(0)=0$.

We prove the case $n=1$ by a direct calculation. Let $X \sim \text{Ber}(p)$ and $Y \sim \text{Poisson}(p)$. Then

$$\begin{aligned}&\sum_{k=0}^{\infty}|P(X=k)-P(Y=k)|\\ &=|P(X=0)-P(Y=0)|+|P(X=1)-P(Y=1)|+\sum_{k=2}^{\infty}P(Y=k)\\ &=|1-p-e^{-p}|+|p-pe^{-p}|+P(Y\geq 2)\\ &=(e^{-p}-(1-p))+(p-pe^{-p})+(1-e^{-p}(1+p))\\ &=p(1-e^{-p})\leq p^2\end{aligned} \tag{9.2}$$

where we applied (9.1) to $x=-p$. The case $n=1$ of Theorem 9.1 has been proved.

For the general case we perform first the following construction. Let $0 \leq p \leq 1$. Define the following discrete probability space: the sample space is $\Omega' = \{-1, 0, 1, \dots\}$ and the probability measure P_p is defined for $k \in \Omega'$ by

$$P_p\{k\}=\begin{cases}1-p, & k=-1\\ e^{-p}-(1-p), & k=0\\ e^{-p}\dfrac{p^k}{k!}, & k\geq 1.\end{cases} \tag{9.3}$$

This is a probability measure because $e^{-p}-(1-p)\geq 0$ by (9.1) and $\sum_{k=0}^{\infty} e^{-p}\frac{p^k}{k!}=1$.

For the proof of the case of general n we construct a coupling of independent Bernoulli variables $X_i \sim \text{Ber}(p_i)$ and independent Poisson variables $Y_i \sim \text{Poisson}(p_i)$. Let the sample space Ω be the n-fold Cartesian product $\{-1, 0, 1, \dots\}^n$. The probability measure P on Ω is defined for $\omega = (\omega_1, \dots, \omega_n) \in \Omega$ by

$$P\{\omega\}=P_{p_1}\{\omega_1\}P_{p_2}\{\omega_2\}\cdots P_{p_n}\{\omega_n\}$$

where each factor $P_{p_i}\{\omega_i\}$ comes from (9.3). This type of P is called a *product probability measure* and denoted by $P = P_{p_1} \otimes P_{p_2} \otimes \cdots \otimes P_{p_n}$.

Let $Z_i(\omega) = \omega_i$ denote the coordinate random variables on Ω. Then for any $k_1, k_2, \dots, k_n \in \{-1, 0, 1, \dots\}$,

$$\begin{aligned} P\{Z_1 = k_1, Z_2 = k_2, \dots, Z_n = k_n\} &= P\{(k_1, k_2, \dots, k_n)\} \\ &= P_{p_1}\{k_1\} P_{p_2}\{k_2\} \cdots P_{p_n}\{k_n\}. \end{aligned}$$

By summing away k_j for $j \neq i$ we conclude from this that $P(Z_i = k_i) = P_{p_i}\{k_i\}$ (P_{p_i} is the probability distribution of Z_i) and then that $Z_1, \dots, Z_n$ are independent.

Define on Ω for $i = 1, 2, \dots, n$ random variables X_i and Y_i as functions of Z_i by

$$X_i = \begin{cases} 0, & Z_i = -1 \\ 1, & Z_i \geq 0 \end{cases} \quad \text{and} \quad Y_i = \begin{cases} 0, & Z_i = -1 \text{ or } 0 \\ Z_i, & Z \geq 1. \end{cases}$$

From the independence of the Z_is follows that $X_1, \dots, X_n$ are independent (among themselves), and $Y_1, \dots, Y_n$ are independent (but X_i and Y_i are not independent). Furthermore, their distributions can be derived from the distributions of the Z_is: $X_i \sim \text{Ber}(p_i)$ and $Y_i \sim \text{Poisson}(p_i)$. From the construction comes the bound

$$\begin{aligned} P(X_i \neq Y_i) &= P(Z_i \notin \{-1, 1\}) = 1 - (1 - p_i + p_i e^{-p_i}) \\ &= p_i(1 - e^{-p_i}) \leq p_i^2. \end{aligned} \tag{9.4}$$

Let $S = \sum_{i=1}^n X_i$ and $Y = \sum_{i=1}^n Y_i$. As a sum of independent Poisson variables, $Y \sim \text{Poisson}(\sum_{i=1}^n p_i)$. We bound First for any given k:

$$\begin{aligned} &|P(S = k) - P(Y = k)| \\ &= |P(S = k, Y = k) + P(S = k, Y \neq k) - P(S = k, Y = k) - P(S \neq k, Y = k)| \\ &= |P(S = k, Y \neq k) - P(S \neq k, Y = k)| \\ &\leq P(S = k, S \neq Y) + P(Y = k, S \neq Y). \end{aligned}$$

Summing this over k gives

$$\begin{aligned} \sum_k |P(S = k) - P(Y = k)| &\leq \sum_k P(S = k, S \neq Y) + \sum_k P(Y = k, S \neq Y) \\ &= 2P(S \neq Y). \end{aligned}$$

Finally we estimate as follows:

$$\begin{aligned} \sum_k |P(S = k) - P(Y = k)| &\leq 2P(S \neq Y) = 2P\Big(\sum_{i=1}^n X_i \neq \sum_{i=1}^n Y_i\Big) \\ &\leq 2P(\text{there is an index } i \in \{1, \dots, n\} \text{ such that } X_i \neq Y_i) \\ &\leq 2\sum_{i=1}^n P(X_i \neq Y_i) \leq 2\sum_{i=1}^n p_i^2. \end{aligned}$$

The last step used (9.4). □

9.2. Total variation distance

Definition 9.3. Suppose that Q_1 and Q_2 are probability distributions of real valued random variables. The **total variation distance** of Q_1 and Q_2 is defined as follows:

$$d_{TV}(Q_1, Q_2) = \sup_{A \subset \mathbb{R}} |Q_1(A) - Q_2(A)| \,. \tag{9.5}$$

The supremum above is over all Borel subsets of $\mathbb{R}$.

The total variation distance $d_{TV}(X_1, X_2)$ of two random variables X_1 and X_2 is by definition the total variation distance $d_{TV}(Q_1, Q_2)$ of their probability distributions Q_1 and Q_2, defined by $Q_i(B) = P(X_i \in B)$ for Borel sets $B \subset \mathbb{R}$. If X_1 and X_2 are defined on the same probability space $(\Omega, \mathcal{F}, P)$, their total variation distance can be calculated on Ω:

$$d_{TV}(X_1, X_2) = \sup_{A \subset \mathbb{R}} |P(X_1 \in A) - P(X_2 \in A)|. \tag{9.6}$$

Theorem 9.4. *If X and Y are integer valued random variables then*

$$d_{TV}(X, Y) = \frac{1}{2} \sum_k |P(X = k) - P(Y = k)| \,.$$

Proof. To prove this, we identify the set on which the supremum in (9.6) is achieved. Let $B = \{k \in \mathbb{Z} : P(X = k) \geq P(Y = k)\}$. Then for any set A, we can bound as follows.

$$\begin{aligned}
&|P(X \in A) - P(Y \in A)| \\
&= \bigl|P(X \in AB) + P(X \in AB^c) - P(Y \in AB) - P(Y \in AB^c)\bigr| \\
&= \Bigl|\sum_{k \in AB} \bigl(P(X = k) - P(Y = k)\bigr) - \sum_{k \in AB^c} \bigl(P(Y = k) - P(X = k)\bigr)\Bigr| \\
&\leq \max\Bigl\{ \sum_{k \in AB} \bigl(P(X = k) - P(Y = k)\bigr) \,,\, \sum_{k \in AB^c} \bigl(P(Y = k) - P(X = k)\bigr)\Bigr\} \\
&\leq \max\Bigl\{ \sum_{k \in B} \bigl(P(X = k) - P(Y = k)\bigr) \,,\, \sum_{k \in B^c} \bigl(P(Y = k) - P(X = k)\bigr)\Bigr\} \\
&= \max\Bigl\{ P(X \in B) - P(Y \in B) \,,\, P(Y \in B^c) - P(X \in B^c)\Bigr\} \\
&= P(X \in B) - P(Y \in B) \\
&\leq d_{TV}(X, Y).
\end{aligned}$$

The first inequality above came from $|x - y| \leq \max\{x, y\}$ which is valid for $x, y \geq 0$. The last equality used

$$\begin{aligned}
P(Y \in B^c) - P(X \in B^c) &= \bigl(1 - P(Y \in B)\bigr) - \bigl(1 - P(X \in B)\bigr) \\
&= P(X \in B) - P(Y \in B).
\end{aligned}$$

Since $d_{TV}(X,Y)$ is the supremum over A of the first line of the calculation, we get the inequalities

$$\begin{aligned} d_{TV}(X,Y) &= \sup_{A\subset\mathbb{R}} |P(X\in A) - P(Y\in A)| \\ &\leq P(X\in B) - P(Y\in B) \leq d_{TV}(X,Y) \end{aligned}$$

which imply

$$d_{TV}(X,Y) = P(X\in B) - P(Y\in B).$$

We can conclude the proof with the following calculation:

$$\begin{aligned} &\tfrac{1}{2}\sum_{k} |P(X=k) - P(Y=k)| \\ &\quad= \tfrac{1}{2}\sum_{k\in B} |P(X=k) - P(Y=k)| + \tfrac{1}{2}\sum_{k\in B^c} |P(X=k) - P(Y=k)| \\ &\quad= \tfrac{1}{2}\big(P(X\in B) - P(X\in B)\big) + \tfrac{1}{2}\big(P(Y\in B^c) - P(X\in B^c)\big) \\ &\quad= P(X\in B) - P(X\in B) = d_{TV}(X,Y). \end{aligned}$$

□

We can now restate Theorem 9.1 in terms of the total variation distance.

Theorem 9.5. *Let $p_1,\dots,p_n\in[0,1]$ and let $X_1,\dots,X_n$ be independent random variables with marginal distributions $X_i\sim\mathrm{Ber}(p_i)$. Let $S=\sum_{i=1}^n X_i$ and $Y\sim\mathrm{Poisson}(\sum_{i=1}^n p_i)$. Then*

$$d_{TV}\Big(\sum_{i=1}^n X_i,\, Y\Big) \leq \sum_{i=1}^n p_i^2.$$

Appendix A

Notation

$\log x$ is the natural logarithm of $x > 0$, also denoted by $\ln x$.

$\mathbb{Z}$ is the set of integers. Restricted sets of integers are indicated by subscripts, as in $\mathbb{Z}_{>0} = \{1, 2, 3, \dots\}$ and $\mathbb{Z}_{\geq b} = \{b, b+1, b+2, \dots\}$ for an integer b.

$\mathbb{R}$ is the set of real numbers. For a positive integer d, $\mathbb{R}^d$ is the *Euclidean space* of real d-vectors:

$$\mathbb{R}^d = \{(x_1, \dots, x_d) : \text{each } x_i \in \mathbb{R}\}.$$

Appendix B

Sets

Let A be a set. The following trichotomy concerning the size of A is important for probability theory.

- A is *finite* if for some positive integer n, there is a bijection $f : \{1, \dots, n\} \to A$. Then n is the *cardinality* of A, in other words, the number of elements in A.
- A is *countably infinite* if there is a bijection $f : \mathbb{Z}_{>0} \to A$. Equivalently, the elements of A can be arranged in a sequence: $A = \{x_1, x_2, x_3, \dots\} = \{x_k\}_{k \in \mathbb{Z}_{>0}}$.
- A is *uncountable* if it is neither finite nor countably infinite.

Finite and countably infinite sets together are referred to as *at most countable* or simply *countable*. Then A is countable if and only if there exists a surjective function $f : \mathbb{Z}_{>0} \to A$, that is, a function f from $\mathbb{Z}_{>0}$ *onto* A. Discussions of countable and uncountable sets can be found in Chapter 2 of [**Rud76**] and Chapter 1 of [**Mun00**]. (Note though that [**Rud76**] uses the term countable as a synonym for countably infinite.)

Examples of countably infinite sets include the set of integers $\mathbb{Z}$ and the set of rationals $\mathbb{Q}$. Any subset of a countable set is countable. Countable unions of countable sets are again countable. Finite Cartesian products of countable sets are again countable.

Any nondegenerate interval of the real line is uncountable. If A is uncountable and $A \subset B$ then B is also uncountable. If S is a set with at least two elements, then the set of S-valued sequences is uncountable. The last fact is proved by the *diagonal argument*. This is such a basic component of a mathematical toolkit that we present it here.

Lemma B.1. *The set of* $\{0, 1\}$*-valued sequences is uncountable.*

Proof. Let A be the set of $\{0,1\}$-valued sequences. Each element $\mathbf{x} \in A$ is a sequence $\mathbf{x} = \{x_1, x_2, x_3, \dots\}$ where each entry x_k is 0 or 1. Suppose that A is countable. Then the elements of A can be arranged in a sequence: $A = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \dots\}$. Since each $\mathbf{x}^j$ is itself a sequence, we write it as $\mathbf{x}^j = \{x_1^j, x_2^j, x_3^j, \dots\}$. (We use boldface $\mathbf{x}$ to distinguish a sequence from its entries x_k, and a superscript j in $\mathbf{x}^j$ to distinguish the indexing of the sequence of elements of A from the indexing of the entries.)

Define a sequence $\mathbf{y} = \{y_k\}$ as follows:

$$y_k = \begin{cases} 0, & \text{if } x_k^k = 1 \\ 1, & \text{if } x_k^k = 0. \end{cases}$$

As a sequence of zeros and ones, $\mathbf{y}$ is an element of A. But $\mathbf{y}$ is not a member of the sequence $\{\mathbf{x}^j\}$ because for each j, $\mathbf{y}$ and $\mathbf{x}^j$ disagree at the jth entry. Thus the set A cannot be captured by a sequence. □

Appendix C

Analysis

Our primary reference for introductory analysis is the classic affectionately known as *Little Rudin* [**Rud76**].

The real number system.

Let $A \subset \mathbb{R}$. A real number b is an *upper bound* of A if $x \leq b$ for all $x \in A$. If A has an upper bound, it is said to be *bounded above.* In that case the *supremum* (or *least upper bound*) of A is the unique real number $c = \sup A$ that satisfies

(i) $x \leq c$ for all $x \in A$; and

(ii) if b is any upper bound of A then $c \leq b$.

That a set with an upper bound has a supremum can be taken as an axiomatic property of the real number system. Or, if the reals are constructed from something more basic, then the existence of a supremum has to be proved.

If the set A has a *maximum* $\max A$, that is, a largest element, then $\sup A = \max A$. The supremum is needed precisely because not every set has a maximum even if it bounded above. For a simple example, take the open unit interval with $\sup(0, 1) = 1$.

If A is not bounded above, then $\sup A = \infty$.

The corresponding term for the greatest lower bound is the *infimum* of A denoted by $\inf A$. If A has a least element, then this is the *minimum* $\min A$ and it coincides with $\inf A$.

The *positive part* and *negative part* of a real number x are

$$x^+ = x \vee 0 \quad \text{and} \quad x^- = (-x) \vee 0. \tag{C.1}$$

$\vee$ is an alternative notation for maximum: $a \vee b = \max\{a, b\}$ = the larger one of a and b. Examples: $5^+ = 5$, $5^- = 0$, $(-3)^+ = 0$, $(-3)^- = 3$. The positive and negative parts $x^\pm$ satisfy $x = x^+ - x^-$ and $|x| = x^+ + x^-$. Examples: $5 = 5^+ - 5^- = 5 - 0$. $|5| = 5^+ + 5^- = 5 + 0$. $-3 = (-3)^+ - (-3)^- = 0 - 3$. $|-3| = (-3)^+ + (-3)^- = 0 + 3$.

Limits of sequences.

For any set S, an S-valued *sequence* is a function from $\mathbb{Z}_{>0}$ into S. The image of an integer k is often denoted by x_k (or with a letter other than x) instead of the usual function notation $f(k)$. Common notations for the entire sequence include $\{x_1, x_2, x_3, \dots\}$, $\{x_k\}_{k\in\mathbb{Z}_{>0}}$ and $(x_k)_{k\in\mathbb{Z}_{>0}}$. The index set can also be left out of the notation: $\{x_k\}$ also denotes the sequence $\{x_k\}_{k\in\mathbb{Z}_{>0}}$. The index set does not have to be $\mathbb{Z}_{>0}$. Other subsets of integers can also appear. In particular, a *bi-infinite* sequence $\{x_k\}_{k\in\mathbb{Z}} = \{\dots, x_{-1}, x_0, x_1, \dots\}$ is one that extends infinitely in both directions.

The *limit superior* ("limsup") of a sequence of numbers $\{x_n\}_{n\ge1}$ is defined as follows, with two alternative notations and two equivalent definitions:

$$\limsup_{n\to\infty} x_n = \varlimsup_{n\to\infty} x_n = \lim_{k\to\infty} \sup_{n:n\ge k} x_n = \inf_{k\in\mathbb{Z}_{>0}} \sup_{n:n\ge k} x_n. \tag{C.2}$$

The last equality is true because the sequence $y_k = \sup_{n:n\ge k} x_n$ is nonincreasing. Thus $\varlimsup x_n$ exists as a number in $[-\infty, \infty]$. Analogously, the *limit inferior* ("liminf") is defined as follows:

$$\liminf_{n\to\infty} x_n = \varliminf_{n\to\infty} x_n = \lim_{k\to\infty} \inf_{n:n\ge k} x_n = \sup_{k\in\mathbb{Z}_{>0}} \inf_{n:n\ge k} x_n. \tag{C.3}$$

The notation can be simplified to $\varlimsup x_n$ and $\varliminf x_n$ when there is no confusion about the index.

The virtue of these concepts is that, while a sequence does not have to have a limit, $\varlimsup_{n\to\infty} x_n$ and $\varliminf_{n\to\infty} x_n$ always exist and they can be investigated as a prelude to proving that a sequence converges. The key fact is that $\{x_n\}$ converges to a value in $[-\infty, \infty]$ if and only if $\varlimsup_{n\to\infty} x_n = \varliminf_{n\to\infty} x_n$, and this case the limit equals $\varlimsup_{n\to\infty} x_n = \varliminf_{n\to\infty} x_n$.

The next lemma collects properties of limsup and liminf. Some statements made only for limsup work for liminf through the identity $\varlimsup(-x_n) = -\varliminf x_n$.

Lemma C.1. *Properties of limsup and liminf. Let $\{x_n\}_{n\ge1}$, $\{y_n\}_{n\ge1}$ and c be numbers in $[-\infty, \infty]$.*

(i) *If $x_n \le c$ for all n, then $\varlimsup x_n \le c$. If $x_n \le y_n$ for all n, then $\varlimsup x_n \le \varlimsup y_n$.*

(ii) *It is always the case that $\varliminf x_n \le \varlimsup x_n$. If $\varlimsup x_n \le c \le \varliminf x_n$ then $c = \lim x_n$.*

(iii) *Let $b \in \mathbb{R}$. Then $b \ge \varlimsup x_n$ if and only if for every $\varepsilon > 0$ there exists $n_0 < \infty$ such that $n \ge n_0$ implies $x_n \le b + \varepsilon$. And $b < \varlimsup x_n$ if and only if there exists $\varepsilon > 0$ such that $x_n \ge b + \varepsilon$ for infinitely many n.*

(iv) *Let A be the set of subsequential limits of $\{x_n\}$:*

$$A = \{x \in [-\infty, \infty] : \exists \text{ subsequence } \{x_{n_j}\}_{j\ge1} \text{ such that } x_{n_j} \to x \text{ as } j \to \infty \}.$$

Then A is a closed subset of $[-\infty, \infty]$, $\varlimsup x_n$ is the maximum of A and $\varliminf x_n$ is the minimum of A.

A fundamental fact is that exponential decay defeats polynomial growth. The next lemma is one manifestation of this.

Lemma C.2. *Fix $k \in \mathbb{Z}_{\geq 0}$ and $\alpha \in (0,1)$, and define the sequence $x_n = n^k \alpha^n$. Then $\lim\limits_{n\to\infty} x_n = 0$.*

Proof. Since

$$\lim_{n\to\infty} \alpha(1+\tfrac{1}{n})^k = \alpha < 1,$$

we can fix a small $\delta > 0$ and n_0 large enough so that $\alpha(1+\frac{1}{n})^k < 1-\delta$ for all $n \geq n_0$. Then, for $n > n_0$,

$$\begin{aligned} x_n &= n^k\alpha^n = \alpha(1+\tfrac{1}{n-1})^k \cdot (n-1)^k\alpha^{n-1} < (1-\delta)x_{n-1} \\ &< (1-\delta)^2 x_{n-2} < \cdots < (1-\delta)^{n-n_0}x_{n_0}. \end{aligned}$$

Keeping n_0 fixed and letting $n \to \infty$, limit (C.12) implies that $x_n \to 0$. □

Infinite series.

Let $\{a_k\}_{k\in\mathbb{Z}_{>0}}$ be a real sequence. The value of the series $\sum_{k=1}^{\infty} a_k$ is by definition the limit of the partial sums $s_n = \sum_{k=1}^{n} a_k$, provided this limit exists:

$$\sum_{k=1}^{\infty} a_k = \lim_{n\to\infty} s_n. \tag{C.4}$$

When the limit exists as a real number, the series is said to *converge*. Note that $\sum_{k=1}^{\infty} a_k$ is not a result of addition, but a limit. The upper summation limit "∞" is symbolic, and does not suggest that there is a last element a_∞. When the range of the summation index is understood, we can abbreviate the notation $\sum_{k=1}^{\infty} a_k$ as $\sum a_k$.

In probability many series of nonnegative terms appear. If $a_k \geq 0$ for all k, then $\{s_n\}$ is a nondecreasing sequence, and it either converges (if the sequence $\{s_n\}$ is bounded) or diverges to infinity (if the sequence $\{s_n\}$ is unbounded). Thus we can always assign a value to a series of nonnegative terms: if the series diverges to infinity, we write $\sum a_k = \infty$. The complementary statement $\sum a_k < \infty$ means that the series converges to a real value.

Series $\sum a_k$ is said to be *absolutely convergent* if $\sum |a_k| < \infty$. For a series of nonnegative terms, absolute convergence is the same as convergence. For all series, absolute convergence implies convergence.

The two important series that appear throughout the book are these. The geometric series the formula to remember is

$$\sum_{n=0}^{\infty} ax^n = \frac{a}{1-x} \quad \text{for any } x \text{ such that } |x| < 1, \tag{C.5}$$

and the series for the exponential function

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}, \qquad \text{for all } x \in \mathbb{R}. \tag{C.6}$$

The geometric series formula is found by first deriving the formula for the partial sums. If

$$s_n = a + ax + ax^2 + \cdots + ax^{n-1} + ax^n$$

then

$$s_n x = ax + ax^2 + ax^3 + \cdots + ax^n + ax^{n+1}$$

and subtraction gives

$$s_n - s_n x = a - ax^{n+1}.$$

If $x \neq 1$, divide by $1 - x$ to get

$$\text{(C.7)} \qquad \sum_{k=0}^{n} ax^k = s_n = \frac{a(1-x^{n+1})}{1-x} \qquad \text{for any } x \neq 1.$$

Formula (C.7) itself is very valuable. If $|x| < 1$, then $x^n \to 0$ as $n \to \infty$ and we can take the limit to obtain (C.5).

The exponential series (C.6) is often taken as the definition of the exponential function. Or, if the exponential function is defined in some other way, then (C.6) is obtained as the Taylor series of the exponential function.

A third useful series is the Taylor series of the function $f(x) = (1+x)^\alpha$ for an arbitrary real α:

$$\text{(C.8)} \qquad (1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n \qquad \text{for } -1 < x < 1.$$

The generalized binomial coefficient is defined by

$$\text{(C.9)} \qquad \binom{\alpha}{0} = 1 \quad \text{and} \quad \binom{\alpha}{n} = \frac{\alpha(\alpha-1)\cdots(\alpha-n+1)}{n!}$$

for $\alpha \in \mathbb{R}$ and $n \in \mathbb{Z}_{>0}$.

A *power series* is of the form $\sum_{n=0}^\infty a_n x^n$. Its *radius of convergence* R is defined by

$$\frac{1}{R} = \varlimsup_{n\to\infty} |a_n|^{1/n}.$$

If $R > 0$, the series converges absolutely for $|x| < R$. This comes as follows. Suppose $0 < |x| < R$ (there is no convergence issue at $x = 0$). Let $\varepsilon > 0$ be such that $|x| < (1-2\varepsilon)R$ which is the same as $1/R < (1-2\varepsilon)/|x|$. Thus

$$\varlimsup_{n\to\infty} |a_n|^{1/n} < \frac{1-2\varepsilon}{|x|}.$$

This implies the existence of n_0 such that $|a_n|^{1/n} < (1-\varepsilon)/|x|$ for $n \geq n_0$. From this, $|a_n x^n| \leq (1-\varepsilon)^n$ for $n \geq n_0$. Since $\sum(1-\varepsilon)^n$ is a convergent geometric series, the series $\sum a_n x^n$ converges absolutely.

If $R > 0$, then inside its radius of convergence the series defines a function $f(x) = \sum_{n=0}^\infty a_n x^n$. A convergent power series can be differentiated terms by term. This identifies the values of the coefficients:

$$a_n = \frac{f^{(n)}(0)}{n!}$$

where $f^{(n)}(0)$ is the nth derivative of f evaluated at zero.

These analytic facts give a useful method for proving that two sequences $\{a_n\}$ and $\{b_n\}$ agree term by term. Namely, if

$$\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n \tag{C.10}$$

and both series converge for all $|x| < R$ for some $R > 0$, then $a_n = b_n$ for all n. When used in this manner, the function $f(x) = \sum_{n=0}^{\infty} a_n x^n$ is called the *generating function* of the sequence $\{a_n\}$.

Functions.

Suppose a function f is defined on an open interval (a, b) on the real line.

The *right limit of f at a* is a real number r if, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $a < x \le a + \delta$ implies $|f(x) - r| \le \varepsilon$. In plain English, $f(x)$ approaches r when x approaches a from the right. In mathematical notation this limit can be expressed as $r = f(a+)$ and $r = \lim\limits_{x \to a+} f(x)$.

The *left limit* at b is denoted by $s = f(b-) = \lim\limits_{x \to b-} f(x)$ and defined analogously: for every $\varepsilon > 0$ there exists $\delta > 0$ such that $b - \delta \le x < b$ implies $|f(x) - s| \le \varepsilon$.

It is convenient that these limits can be accessed with monotone sequences, as stated in the next lemma. A sequence $\{a_n\}$ in $\mathbb{R}$ is *strictly increasing* if $a_n < a_{n+1}$ for each n, and *strictly decreasing* if $a_n > a_{n+1}$ for each n.

Lemma C.3.

(a) *$r = f(a+)$ if and only if $f(x_n) \to r$ for every sequence $x_n \in (a, b)$ such that x_n is strictly decreasing and $x_n \to a$.*

(b) *$s = f(b-)$ if and only if $f(y_n) \to s$ for every sequence $y_n \in (a, b)$ such that y_n is strictly increasing and $y_n \to b$.*

Proof. We address part of (a). The implication from $r = f(a+)$ to $f(x_n) \to r$ for every sequence x_n that decreases strictly to a is straightforward. We prove the converse (opposite) implication by verifying its *contrapositive*.[1] Suppose $r = f(a+)$ fails. Negating the definition of $r = f(a+)$ given above gives this statement:

$$\begin{aligned} &\text{there exists } \varepsilon > 0 \text{ such that for every } \delta > 0 \\ &\text{there exists } x \in (a, a+\delta) \text{ such that } |f(x) - r| > \varepsilon. \end{aligned} \tag{C.11}$$

We apply this statement inductively to construct a sequence $\{x_n\}$.

- Step 1: apply the statement to $\delta = 1$ to choose $x_1 \in (a, a+1)$ such that $|f(x_1) - r| > \varepsilon$.
- Let $n \in \mathbb{Z}_{>0}$. Assume inductively that we have constructed $x_1 > x_2 > \ldots > x_n > a$ such that $x_k \in (a, a + \frac{1}{k})$ and $|f(x_k) - r| > \varepsilon$ for all $k = 1, \ldots, n$. To go to step $n+1$, apply (C.11) to $\delta = \frac{1}{n+1} \wedge \frac{1}{2}(x_n - a)$ to pick a point $x_{n+1} \in (a, a+\delta)$ such that $|f(x_{n+1}) - r| > \varepsilon$. Now the claim has been

[1] The contrapositive of P *implies* Q is *not-Q implies not-P*. The two statements are logically equivalent, that is, they are either both true or both false. Thus either one can be proved by proving the other.

extended to $n+1$: we have $x_1 > x_2 > \ldots > x_n > x_{n+1} > a$ such that $x_k \in (a, a+\frac{1}{k})$ and $|f(x_k) - r| > \varepsilon$ for all $k = 1, \ldots, n+1$.

This inductive construction produces an infinite sequence $\{x_n\}$ that decreases strictly to a while $|f(x_n) - r| > \varepsilon$ for all n. In particular, this sequence fails the condition $f(x_n) \to r$.

To summarize, if $r = f(a+)$ fails then there exists a strictly decreasing sequence $x_n \to a$ that fails the condition $f(x_n) \to r$. This establishes the statement we want: namely that, if $f(x_n) \to r$ for every strictly decreasing sequence $x_n \to a$, then $r = f(a+)$. □

If f is defined on an interval $[a, b)$, then f is *right-continuous at a* if $f(a) = f(a+)$. Similarly, if f is defined on an interval $(a, b]$, then f is *left-continuous at b* if $f(b) = f(b-)$. If f is defined on an open interval (a, b) and $x \in (a, b)$, then *continuity* of f at x is equivalent to $f(x-) = f(x) = f(x+)$.

Definition C.4. Let f be a function defined on $\mathbb{R}$ or some subinterval of $\mathbb{R}$. Then we say that f is **piecewise continuous** if for every bounded open subinterval (a, b) of the domain of f there exists a finite partition $a = s_0 < s_1 < \cdots < s_m = b$ such that f is continuous on each open partition interval (s_{i-1}, s_i), $i = 1, \ldots, m$, and one-sided limits $f(s_0+)$, $f(s_i\pm)$ for $i = 1, \ldots, m-1$, and $f(s_m-)$ exist. △

We take for granted the following two basic limits:

$$\lim_{x\to\infty} \rho^x = 0 \qquad \text{for all } 0 < \rho < 1 \tag{C.12}$$

and

$$\lim_{x\to\infty} \Big(1 + \frac{a}{x}\Big)^x = e^a \qquad \text{for all } a \in \mathbb{R}. \tag{C.13}$$

We think of the limits above as limits of functions, so we wrote $x \to \infty$ to indicate that the variable x tends to infinity along the reals. A special case is the limit where x tends to infinity along some sequence. For example, if x tends to infinity along integers we would write n in place of x.

Integration.

As measure theory is not a prerequisite for this book, we use the Riemann integral studied in calculus and introductory analysis. For a bounded real-valued function f on a compact interval $[a, b]$, the *Riemann integral* $\int_a^b f(x)\,dx$ is defined by the limit

$$\int_a^b f(x)\,dx = \lim_{\text{mesh}(\mathcal{P})\to 0} \sum_{i=1}^n f(s_i)\,\Delta x_i \tag{C.14}$$

where $\mathcal{P} = \{a = x_0 < x_2 < \cdots < x_n = b\}$ is a *partition* of the interval $[a, b]$, $\Delta x_i = x_i - x_{i-1}$ is the length of the ith partition interval, $s_i \in [x_{i-1}, x_i]$ are arbitrary choices of points from the partition intervals, and $\text{mesh}(\mathcal{P}) = \max \Delta x_i$ is the maximal length of a partition interval. The approximating sums $\sum_{i=1}^n f(s_i)\,\Delta x_i$ are called *Riemann sums*. The integration variable x does not add any information to the notation $\int_a^b f(x)\,dx$, so we can readily simplify $\int_a^b f(x)\,dx$ to $\int_a^b f$.

The precise meaning of definition (C.14) is that, in order for the integral to exist and have value $c = \int_a^b f$, it is required that for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\left| c - \sum_{i=1}^{n} f(s_i)\,\Delta x_i \right| < \varepsilon \tag{C.15}$$

for all partitions $\mathcal{P}$ with $\text{mesh}(\mathcal{P}) < \delta$ and for any choice of points $s_i \in [x_{i-1}, x_i]$.

A basic theorem states that if a bounded function f is continuous on $[a, b]$, except possibly at finitely many points, the integral $\int_a^b f$ exists. In particular, the piecewise continuous functions of Definition C.4 can be integrated. The definition (C.14) is rarely used to evaluate an integral. Calculus is used for that purpose.

Improper integrals are defined as limits: for example

$$\int_a^\infty f = \lim_{b\to\infty} \int_a^b f \quad \text{and} \quad \int_{-\infty}^\infty f = \lim_{a\to-\infty,\, b\to\infty} \int_a^b f\,. \tag{C.16}$$

The meaning of the last limit is that $c = \int_{-\infty}^\infty f$ holds if for every $\varepsilon > 0$ there exists $M > 0$ such that

$$\left| c - \int_a^b f \right| < \varepsilon \qquad \text{for all } a < -M \text{ and } b > M.$$

Equivalently, for *all* sequences $a_n \to -\infty$ and $b_n \to \infty$, $\int_{a_n}^{b_n} f \to c$ as $n \to \infty$. Analogous statements hold for the first limit of (C.16).

When $f \geq 0$ on the interval of integration, either the limits in (C.16) exist as real numbers, or the sequences in question diverge to infinity. In the latter case we write (in the case of the second statement of (C.16)) $\int_{-\infty}^\infty f = \infty$. Thus for $f \geq 0$, improper integrals always have a well-defined value, as long as the integrals of f over bounded intervals are defined.

Appendix D

Asymptotics

Asymptotics for the logarithm.

We begin with estimates for the natural logarithm function. We write $\log x$ instead of $\ln x$. In the first lemma below we establish its Taylor series with an error term.

Lemma D.1. *Let $0 < \delta < 1$. Then for all $x \in [-1+\delta, 1-\delta]$ and $n \in \mathbb{Z}_{>0}$ we have this estimate:*

$$\left|\log(1+x) - \left(x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots + \frac{(-1)^{n-1}x^n}{n}\right)\right| \le \frac{|x|^{n+1}}{(n+1)\delta} \tag{D.1}$$

Proof. Let $-1+\delta \le s \le 1-\delta$. From

$$\frac{1}{1+s} = \sum_{k=0}^{\infty}(-s)^k$$

we derive for $n \ge 1$

$$\left|\frac{1}{1+s} - \sum_{k=0}^{n-1}(-s)^k\right| = \left|\sum_{k=n}^{\infty}(-s)^k\right| \le \sum_{k=n}^{\infty}|s|^k = \frac{|s|^n}{1-|s|} \le \frac{|s|^n}{\delta}.$$

Next an integration step. At the end we want a nonnegative bound. Due to the sign change $\int_0^x = -\int_x^0$ in a Riemann integral that comes from the order of the real line, we estimate separately for $x \ge 0$ and $x \le 0$. Consider first the case $0 \le x \le 1-\delta$.

$$\left|\log(1+x) - \sum_{k=1}^{n}(-1)^{k-1}\frac{x^k}{k}\right| = \left|\int_0^x \frac{1}{1+s}\,ds - \sum_{k=1}^{n}(-1)^{k-1}\int_0^x s^{k-1}\,dx\right|$$
$$\le \int_0^x \left|\frac{1}{1+s} - \sum_{k=0}^{n-1}(-s)^k\right| ds \le \frac{1}{\delta}\int_0^x s^n\,ds = \frac{x^{n+1}}{(n+1)\delta}.$$

The case $-1+\delta \le x \le 0$ goes similarly. At the end we have to distinguish between even and odd n to get the sign right.

$$\Big|\log(1+x) - \sum_{k=1}^{n}(-1)^{k-1}\frac{x^k}{k}\Big| = \Big| -\int_x^0 \frac{1}{1+s}\,ds + \sum_{k=1}^{n}(-1)^{k-1}\int_x^0 s^{k-1}\,dx\Big|$$
$$\le \int_x^0 \Big|\frac{1}{1+s} - \sum_{k=0}^{n-1}(-s)^k\Big|\,ds \le \frac{1}{\delta}\int_x^0 |s|^n\,ds$$
$$= \begin{cases} \dfrac{1}{\delta}\displaystyle\int_x^0 s^n\,ds = -\dfrac{x^{n+1}}{(n+1)\delta} = \dfrac{|x|^{n+1}}{(n+1)\delta} & \text{if } n \text{ is even,} \\ -\dfrac{1}{\delta}\displaystyle\int_x^0 s^n\,ds = \dfrac{x^{n+1}}{(n+1)\delta} = \dfrac{|x|^{n+1}}{(n+1)\delta} & \text{if } n \text{ is odd.} \end{cases}$$

The lemma is proved. □

By letting $n \to \infty$ in (D.1), we get the series expansion

$$\log(1+x) = \sum_{k=1}^{\infty}(-1)^{k-1}\frac{x^k}{k} \qquad \text{valid for } -1 < x < 1. \tag{D.2}$$

From this we get conveniently the case $n = 0$ of estimate (D.1), with $0 < \delta < 1$:

$$\big|\log(1+x)\big| \le \sum_{k=1}^{\infty}|x|^k = \frac{|x|}{1-|x|} \le \frac{|x|}{\delta} \qquad \text{for } -1+\delta < x < 1-\delta. \tag{D.3}$$

For negative x all terms in the series (D.2) are negative, and we have the following bound:

$$\log(1-t) \le -t - \frac{t^2}{2} \qquad \text{for } 0 < t < 1. \tag{D.4}$$

The next lemma gives two estimates beyond the validity of the series.

Lemma D.2. *We have these bounds for the natural logarithm:*

$$\log(1+x) \le x - \frac{x^2}{4} \qquad \textit{for } 0 \le x \le 1 \tag{D.5}$$

and

$$\log(1+x) \le \frac{x}{2} \qquad \textit{for } x \ge 1. \tag{D.6}$$

Proof. For (D.5), let $g(x) = x - \frac{x^2}{4} - \log(1+x)$. Then $g(0) = 0$ and

$$g'(x) = 1 - \frac{x}{2} - \frac{1}{1+x} = \frac{x - x^2}{2(1+x)} > 0 \quad \text{for } 0 < x < 1.$$

For (D.6), let $h(x) = \frac{x}{2} - \log(1+x)$. Then $h(0) = 0$ and

$$h'(x) = \frac{1}{2} - \frac{1}{1+x} > 0 \quad \text{for } x > 1.$$ □

The big-oh notation means the following. To write $f(x) = O(|x|^m)$ as $x \to 0$ means that there exists a constant C such that $|f(c)| \le C|x|^m$ for all small enough x. For example, from (D.1) we get this estimate that will be used repeatedly in the proof of the CLT:

$$\log(1+x) = x - \frac{x^2}{2} + O(|x|^3) \qquad \text{as } x \to 0. \tag{D.7}$$

Next we state and prove a form of Stirling's formula that gives asymptotics for the factorial $n!$ as n gets large. We prove the result with an approach from asymptotic analysis known as *Laplace's method*. First we explain informally the idea behind Laplace's method, without being technically precise.

Laplace's method.

The goal of Laplace's method is to understand the asymptotics of an integral of the type $\int_a^b e^{nf(x)}\,dx$ as $n \to \infty$. Suppose f has a unique strict maximum at x_0 in the interval (a,b). Then $f'(x_0) = 0$. Assume that $f''(x_0) < 0$. Then around x_0 we can Taylor expand f as

$$f(x) = f(x_0) + \tfrac{1}{2}f''(x_0)(x-x_0)^2 + \text{ small error.} \tag{D.8}$$

Next write the integral as

$$\int_a^b e^{nf(x)}\,dx = e^{nf(x_0)} \int_a^b e^{n(f(x)-f(x_0))}\,dx.$$

Since $f(x) - f(x_0) < 0$ at all $x \ne x_0$, it is useful to separate the integral into the part close to x_0 and the rest. The rest should not amount to much because $e^{nc} \to 0$ fast for negative c. So let $\varepsilon > 0$ be small and decompose the integral into three pieces:

(D.9)

$$\begin{aligned}
&\int_a^b e^{nf(x)}\,dx = e^{nf(x_0)} \int_a^b e^{n(f(x)-f(x_0))}\,dx \\
&= e^{nf(x_0)}\biggl(\int_a^{x_0-\varepsilon} e^{n(f(x)-f(x_0))}\,dx + \int_{x_0-\varepsilon}^{x_0+\varepsilon} e^{n(f(x)-f(x_0))}\,dx + \int_{x_0+\varepsilon}^{b} e^{n(f(x)-f(x_0))}\,dx \biggr) \\
&\approx e^{nf(x_0)}\biggl(2e^{-n\delta} + \int_{x_0-\varepsilon}^{x_0+\varepsilon} e^{-\frac{1}{2}n|f''(x_0)|(x-x_0)^2}\,dx \biggr)
\end{aligned}$$

In the last (imprecise) approximation step we supposed that $f(x)-f(x_0) < -\delta < 0$ for $x \notin (x_0-\varepsilon, x_0+\varepsilon)$, and applied expansion (D.8) but ignored the error term. Next ignore the small exponential $e^{-n\delta}$ and change variables in the remaining integral to $y = \sqrt{n|f''(x_0)|}\,(x-x_0)$. The approximation above becomes

$$\begin{aligned}
\int_a^b e^{nf(x)}\,dx &\approx e^{nf(x_0)} \frac{1}{\sqrt{n|f''(x_0)|}} \int_{-\varepsilon\sqrt{n|f''(x_0)|}}^{\varepsilon\sqrt{n|f''(x_0)|}} e^{-y^2/2}\,dy \\
&\approx \frac{e^{nf(x_0)}}{\sqrt{n|f''(x_0)|}} \int_{-\infty}^{\infty} e^{y^2/2}\,dy = \sqrt{\frac{2\pi}{n|f''(x_0)|}}\; e^{nf(x_0)}.
\end{aligned} \tag{D.10}$$

In the second approximation above we extended the Gaussian integral to the entire real line, justifed by the smallness of the extreme tails of the Gaussian. When the approximations made along the way can be justified, in the end we have the limit

$$\lim_{n\to\infty} \sqrt{n}\, e^{-nf(x_0)} \int_a^b e^{nf(x)}\, dx = \sqrt{\frac{2\pi}{|f''(x_0)|}}\,. \tag{D.11}$$

In the proof below we execute this strategy with precise details.

Stirling's formula.

Theorem D.3 (Stirling's formula)**.**

$$\lim_{n\to\infty} \frac{n!}{n^n e^{-n}\sqrt{2\pi n}} = 1. \tag{D.12}$$

Proof. To use Laplace's method we turn $n!$ into the form $\int_a^b e^{nf(x)}\, dx$. The starting point is the integral formula for $n!$ given by the definition of $\Gamma(n+1)$, proved by integration by parts. A change of variables reveals that the relevant function is $f(x) = \log x - x$ with a unique maximum $f(1) = -1$. We separate this maximum outside the integral as done above in (D.9).

$$\begin{aligned} n! &= \int_0^\infty s^n e^{-s}\, ds = n^{n+1} \int_0^\infty x^n e^{-nx}\, dx = n^{n+1} \int_0^\infty e^{n(\log x - x)}\, dx \\ &= n^{n+1} e^{-n} \int_0^\infty e^{n(\log x - x + 1)}\, dx. \end{aligned}$$

We have the following representation of the ratio in (D.12), except for the constant factor:

$$\frac{n!}{n^n e^{-n}\sqrt{n}} = n^{1/2} \int_0^\infty e^{n(\log x - x + 1)}\, dx. \tag{D.13}$$

Knowing that the maximum is at $x = 1$, we decompose the integral into three parts. We choose an interval of radius n^{-a} around $x = 1$ with $1/3 < a < 1/2$. This choice cannot be seen as the correct one at this early stage of the proof. It comes from hindsight, after observing how the calculation unfolds.

(D.14)

$$\begin{aligned} n^{1/2} \int_0^\infty e^{n(\log x - x + 1)}\, dx \;&=\; n^{1/2} \int_0^{1-n^{-a}} e^{n(\log x - x + 1)}\, dx \\ &\quad + n^{1/2} \int_{1-n^{-a}}^{1+n^{-a}} e^{n(\log x - x + 1)}\, dx + n^{1/2} \int_{1+n^{-a}}^\infty e^{n(\log x - x + 1)}\, dx. \end{aligned}$$

The first and last term on the right are error terms that vanish in the limit, and the nontrivial contribution comes from the middle term. We treat the terms in turn, starting with the middle one.

Middle term on the right of (D.14). Below we use the expansion

$$\log(1+t) = t - \frac{t^2}{2} + O(|t|^3)$$

from (D.1) and then observe that for $|t| \le n^{-a}$, $O(|t|^3) = O(n^{-3a})$. Change of variable $t = s/\sqrt{n}$ then takes us to an integral of the standard normal density.

$$
\begin{aligned}
n^{1/2} \int_{1-n^{-a}}^{1+n^{-a}} e^{n(\log x - x + 1)}\, dx &= n^{1/2} \int_{-n^{-a}}^{n^{-a}} e^{n(\log(1+t)-t)}\, dt \\
&= n^{1/2} \int_{-n^{-a}}^{n^{-a}} e^{-\frac{1}{2}nt^2 + O(n^{1-3a})}\, dt = e^{O(n^{1-3a})} \int_{-n^{\frac{1}{2}-a}}^{n^{\frac{1}{2}-a}} e^{-s^2/2}\, ds
\end{aligned}
$$

Precisely speaking, the calculation above gives these bounds, for some positive constant C:

$$
\begin{aligned}
e^{-Cn^{1-3a}} \int_{-n^{\frac{1}{2}-a}}^{n^{\frac{1}{2}-a}} e^{-s^2/2}\, ds \;&\le\; n^{1/2} \int_{1-n^{-a}}^{1+n^{-a}} e^{n(\log x - x + 1)}\, dx \\
&\le\; e^{Cn^{1-3a}} \int_{-n^{\frac{1}{2}-a}}^{n^{\frac{1}{2}-a}} e^{-s^2/2}\, ds.
\end{aligned}
$$

Assumption $1/3 < a < 1/2$ gives the convergence $e^{\pm Cn^{1-3a}} \to 1$ as $n \to \infty$. We get the limit

$$
\begin{aligned}
\lim_{n\to\infty} n^{1/2} \int_{1-n^{-a}}^{1+n^{-a}} e^{n(\log x - x + 1)}\, dx &= \lim_{n\to\infty} \int_{-n^{\frac{1}{2}-a}}^{n^{\frac{1}{2}-a}} e^{-s^2/2}\, ds \\
&= \int_{-\infty}^{\infty} e^{-s^2/2}\, ds = \sqrt{2\pi}.
\end{aligned}
\tag{D.15}
$$

This is the contribution we want. It remains to show that the first and third term on the right of (D.14) converge to zero as $n \to \infty$.

First term on the right of (D.14)*.* For $0 < x < 1$ first use the fact that the function $f(x) = \log x - x + 1$ is strictly increasing and then the bound $\log(1-x) \le -x - x^2/2$.

$$
\begin{aligned}
n^{1/2} \int_0^{1-n^{-a}} e^{n(\log x - x + 1)}\, dx &\le n^{1/2} e^{n(\log(1-n^{-a}) + n^{-a})} \\
&\le n^{1/2} e^{-\frac{1}{2}n^{1-2a}} \to 0 \quad \text{as } n \to \infty.
\end{aligned}
\tag{D.16}
$$

Third term on the right of (D.14)*.* Change variables, split the integral into two and apply (D.5) and (D.6).

$$
\begin{aligned}
n^{1/2} \int_{1+n^{-a}}^{\infty} e^{n(\log x - x + 1)}\, dx &= n^{1/2} \int_{n^{-a}}^{\infty} e^{n(\log(1+t)-t)}\, dt \\
&= n^{1/2} \int_{n^{-a}}^{1} e^{n(\log(1+t)-t)}\, dt + n^{1/2} \int_{1}^{\infty} e^{n(\log(1+t)-t)}\, dt \\
&\le n^{1/2} \int_{n^{-a}}^{1} e^{-nt^2/4}\, dt + n^{1/2} \int_{1}^{\infty} e^{-nt/2}\, dt \\
&\le n^{1/2} e^{-\frac{1}{4}n^{1-2a}} + n^{1/2} \cdot \frac{2}{n} e^{-n/2} \to 0 \quad \text{as } n \to \infty.
\end{aligned}
\tag{D.17}
$$

The limit (D.12) now comes by substituting limits (D.15), (D.16) and (D.17) into (D.14). □

There is a technical notation for limits such as (D.12). Namely,

$$a_n \sim b_n \quad \text{means that} \quad \frac{a_n}{b_n} \to 1 \quad \text{as } n \to \infty. \tag{D.18}$$

In these terms Stirling's formula says that $n! \sim n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$.

Appendix E

Table of values for $\Phi(x)$

$\Phi(x) = P(Z \leq x)$ is the cumulative distribution function of the standard normal random variable Z.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Answers to selected exercises

Chapter 1.

2.1. (a) $\frac{1}{3}$. (b) $\frac{29}{63} \approx 0.460$. (c) $\frac{4}{9} \approx 0.444$.

2.2. (a) $p_X(1) = \frac{3}{5}$, $p_X(2) = \frac{3}{10}$, $p_X(3) = \frac{1}{10}$.

2.5. $p_Y(k) = \frac{1}{36}(13 - 2k)$ for $k \in \{1, \dots, 6\}$.

2.29. $p_{X_1,Y}(k,k) = \frac{1}{36}(7 - k)$ for $k \in \{1, \dots, 6\}$, and $p_{X_1,Y}(k,\ell) = \frac{1}{36}$ for $k > \ell$ in $\{1, \dots, 6\}$.

Chapter 2.

1.2. (b) $P(A_1) = 1 - \left(\frac{2}{3}\right)^n$ and $P(A_2) = 1 - \left(\frac{2}{3}\right)^n - \frac{1}{2}n\left(\frac{2}{3}\right)^n$.

(c) $P(C_n) = \frac{2^n - 1}{3^{n-1}}$.

(d) $P(C) = 0$.

(e) $P(D) = \frac{1}{2}$.

(f) $P(U) = \frac{2}{5}$.

(g) $P(B_{m,n}) = \left(\frac{2}{3}\right)^{n-m}$, $P(B_{m,\infty}) = 0$, and $P(E) = 1$.

1.3. (b) $P(A) = \frac{2}{5}$. (c) $P(B) = \frac{9}{10}$. (d) $P(C) = \frac{1}{6}$.

1.4. (b) $P(A) = \frac{2}{7}$. (c) $P(A) = \frac{1}{3}$.

1.6. $P(B) = \frac{13}{1015} \approx 0.0128$.

1.7. $\frac{1}{4}$.

1.8. (b) $\frac{3}{4} \le P(A \cup B) \le 1$.

1.16. (a) 0.48, (b) 0.48, (c) 0.12.

1.19. $P(A) = \frac{\frac{5}{6} \cdot (1-\theta)}{1 - \frac{5}{6}\theta}$.

1.20. 0.5.

1.22. (b) 2/3. (c) 1/2.

1.31. 931/1000.

1.33. (a) After one flip, 2/17. After two flips, 4/29. (b) 25.

1.34. (a) $\frac{1}{3}$. (b) $\frac{103}{300} \approx 0.3433$. (c) $\frac{99}{103} \approx 0.9612$.

Chapter 3.

3.1. The joint probability mass function is $P(N = n, Y = x) = \frac{m-n}{m(m-1)}$ for $n \in \{1, \dots, m-1\}$ and $x \in \{a, b\}$.

3.7. $\frac{38{,}400}{43{,}046{,}721} \approx 0.00089$.

3.10. (a) Tails is seen at least once after the mth flip with probability $\frac{1}{2}$. (b) Each heads has marginal probability 0.45. Heads is seen at least once after the mth flip with probability 0.9.

3.11. For $m \in \mathbb{Z}_{>0}$, $P(K = m) = \begin{cases} p(1-p)\frac{p^m-(1-p)^m}{2p-1}, & p \neq \frac{1}{2} \\ m2^{-m-1}, & p = \frac{1}{2}. \end{cases}$

3.12. (a) $(1-p)^2 p$. (b) $p(1-p)$.

3.14. (b) $(\frac{2}{3})^{k-1}$.

3.15. $\frac{\binom{78}{53}}{\binom{80}{55}}$.

3.18. (a) $e^{-6}6^k/k!$. (b) $F \sim \text{Exp}(3)$.

3.22. (a) $\frac{pr}{p+r-pr}$.
(b) $P(Z = n) = pr\frac{(1-r)^{n-1}-(1-p)^{n-1}}{p-r}$ for $n \geq 2$.

3.26.

$$p_{X+Y}(a) = \begin{cases} \frac{a-1}{mn} & 2 \leq a \leq n \\ \frac{1}{m} & n+1 \leq a \leq m+1 \\ \frac{m+n+1-a}{mn} & m+2 \leq a \leq m+n \end{cases}$$

3.27. (a) $P(Y - X \geq \frac{3}{2}) = \frac{1}{24}$.
(b) $f_{X+Y}(z) = \begin{cases} (z-1)^2 & 1 < z < 2 \\ 1-(z-2)^2 & 2 < z < 3 \end{cases}$.

3.28. $f_{X+Y}(z) = \begin{cases} \frac{z-8}{2} & 8 \leq z < 9 \\ \frac{1}{2} & 9 \leq z < 10 \\ \frac{11-z}{2} & 10 \leq z \leq 11 \\ 0 & \text{otherwise.} \end{cases}$

3.29. $f_{X+Y}(z) = \begin{cases} \lambda\mu\frac{e^{-\lambda z}-e^{-\mu z}}{\mu-\lambda}, & \text{if } z > 0 \\ 0, & \text{otherwise.} \end{cases}$

3.31. (a) $\frac{1}{2}e^{-4}$.
(b) $f_{X-Y}(z) = \frac{1}{2}e^{-|z|}$.

3.42. $r_x = \dfrac{1-(\frac{1-p}{p})^x}{1-(\frac{1-p}{p})^M}$.

3.43. $P(S_n = 0 \text{ for some } n) = 1$.

Chapter 4.

4.3. (a) $E[X+Y] = \frac{1}{p} + nr$. (b) $E[X^2+Y^2] = \frac{2-p}{p^2} + n(n-1)r^2 + nr$. (c) Cannot be calculated.

4.9. $EX = 1/p$ and $\operatorname{Var}(X) = (1-p)/p^2$.

4.10. $EX = \frac{3}{5}$ and $\operatorname{Var}(X) = \frac{66}{25}$.

4.11. $EX = \lambda^{-1}$ and $\operatorname{Var}(X) = \lambda^{-2}$.

4.12. $EY = 3$. $EX = \frac{15}{8}$.

4.14. (a) $\operatorname{Cov}(X,Y) - \frac{1}{144}$. (b) $\operatorname{Var}(U) = \frac{3}{80}$.

4.15. $E[X] = \frac{781}{256}$. $\operatorname{Var}(X) \approx 0.4232$.

4.16. $E[X] = 6\left(1-\left(\frac{5}{6}\right)^4\right)$, $\operatorname{Var}(X) \approx 0.447$.

4.17. (a) $\frac{738}{125}$.
(b) $\operatorname{Var}(X) \approx 0.8092$.

4.18. $E[X] = \frac{51}{8}$, $\operatorname{Var}(X) = \frac{1955}{832}$.

Chapter 5.

Chapter 6.

6.7. There is a weak limit if and only if $\sigma \in [0,\infty)$.

6.9. 0.6826 without continuity correction.

6.11. 2401

6.12. (0.426,0.488)

6.13. $P(X \geq 48) \approx 0.1056$. $P(Y \geq 2) \approx 0.2642$.

Chapter 7.

7.1. (a) $M_Z(t) = \begin{cases} 1, & t = 0 \\ \dfrac{e^t - 1}{t}, & t \neq 0. \end{cases}$

7.2. (a) $M_X(t) = \begin{cases} \left(\dfrac{\lambda}{\lambda - t}\right)^r, & t < \lambda \\ \infty, & t \geq \lambda. \end{cases}$

7.3. (a) $M_X(t) = \begin{cases} \dfrac{pe^t}{1-e^t(1-p)}, & t < \log \frac{1}{1-p} \\ \infty, & t \geq \log \frac{1}{1-p}. \end{cases}$

7.4. (a) $f_Y(t) = \begin{cases} \frac{1}{\sqrt{2\pi t^2}} \exp\left(-\frac{(\log t)^2}{2}\right), & t > 0 \\ 0, & t \leq 0. \end{cases}.$

Chapter 8.

8.1. (a) $p_{Y|X}(y|x) = \binom{y-1}{x-1}\left(\frac{1}{2}\right)^y$, for $1 \leq x \leq y$.

8.3. For $0 \leq \ell \leq n$,

$$p_L(\ell) = \binom{n}{\ell}(pr)^\ell(1-pr)^{n-\ell}.$$

8.4. (b) Given $X = x$, $Y \sim \text{Gamma}(2, x+1)$.

8.5. (a) $f_Y(y) = e^{-y}$ for $y > 0$. $f_{X|Y}(x|y) = \frac{1}{y}e^{-x/y}$ for $x > 0$ and $y > 0$. $E[Y] = 1$.

(b) $E[X|Y] = Y$.

(c) $E[X] = E[Y] = 1$.

8.7. $f_Z(z) = \frac{1}{2}(\ln z)^2$ for $0 < z < 1$, $E(Z) = \frac{1}{8}$, and $\text{Var}(Z) = \frac{37}{1728} \approx 0.021$.

Chapter 9.

Bibliography

[Coh80] Donald L. Cohn, Measure theory, Birkhäuser, Boston, Mass., 1980. MR 578344

[Dur10] Rick Durrett, Probability: theory and examples, fourth ed., Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2010. MR 2722836 (2011e:60001)

[Fol99] Gerald B. Folland, Real analysis: Modern techniques and their applications, second ed., Pure and Applied Mathematics, John Wiley & Sons Inc., New York, 1999. MR 2000c:00001

[Mun00] James R. Munkres, Topology, second ed., Prentice-Hall, Inc., Upper Saddle River, N.J., 2000.

[Rud76] Walter Rudin, Principles of mathematical analysis, third ed., McGraw-Hill Publishing Co., 1976.

Index