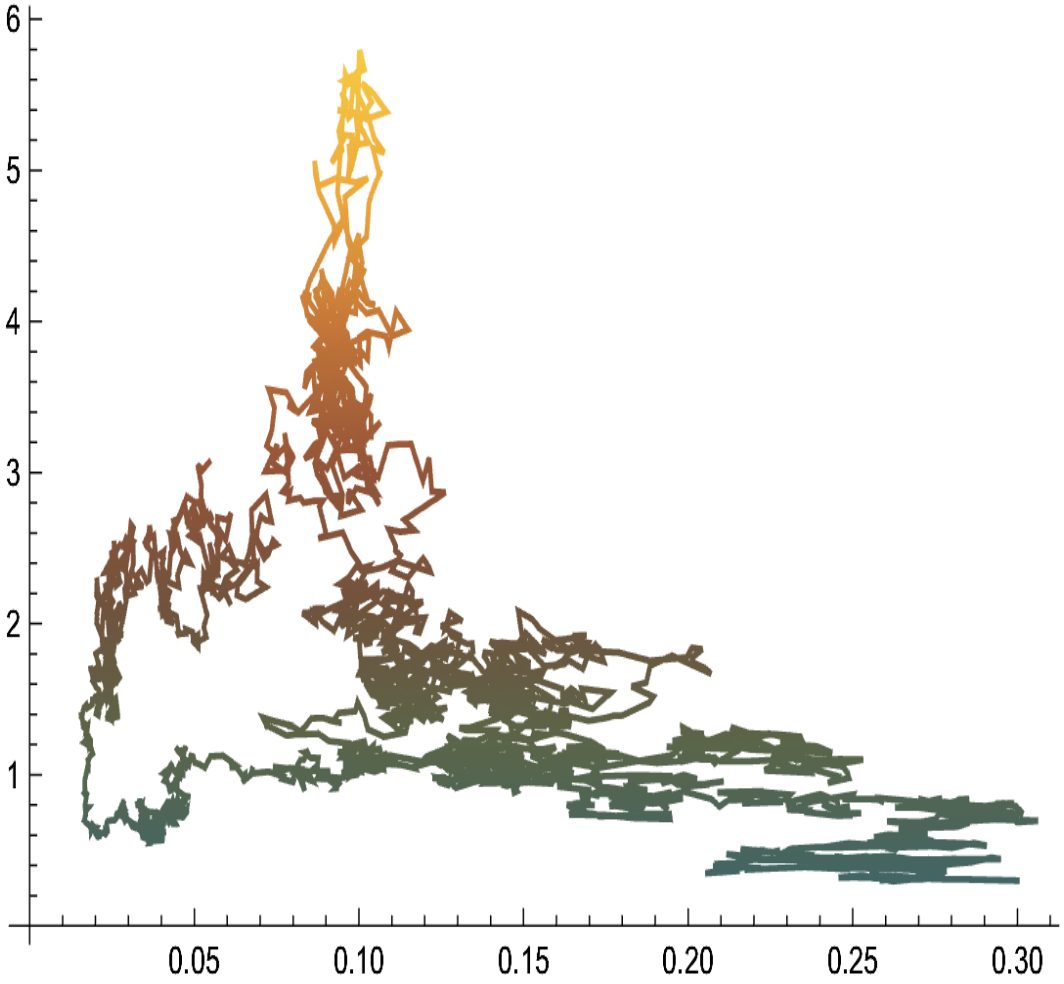


Introductory Probability

Evan Randles

Draft date March 5, 2024

Contents

Preface

These are the course notes for Probability (MA 381). Many of the ideas presented in these notes are not original and, at least in part, have been influenced by a number of excellent texts on the subject, including *Elementary Probability* by Kai Lai Chung and Farid AitSahlia [14], *Probability and Random Processes* by G. R. Grimmett and D. R. Stirzaker [9], and *A First Course in Probability* by Sheldon Ross [11]. As these notes represent an active working draft, please update/download them frequently as I will often make corrections and changes without explicit warning. Any block of text or word in red is just a note for me (one I'm leaving to myself while editing); these should be disregarded. Words in blue are often hyperlinks to an interesting reference, usually a video, and you should click on them if you're reading along digitally. Also, if you find or suspect an error or typo – no matter how trivial – please email me to let me know!

Acknowledgment:

These course notes are a growing document and are, at present, pretty rough. I owe a great deal of gratitude to my students, who have served as test subjects for these notes as well as editors and proof readers. In particular, I would like to thank Andy Day, Selina He, Dean Hickman, Saki Imai, Charlie Ma, Meredith Marra, Iris Liu, Lila Reznick, Gabby Rickards, Yiheng Su, and Ziyan (Sophie) Zhang. Of course, I am responsible for the errors that inevitably remain.

Chapter 1

Experiments: Sample Spaces and Events

Probability is concerned with describing, computing, and predicting the likelihood of outcomes of experiments. For instance, you can think of a game of cards as an experiment where players draw cards; the outcomes are the types of cards drawn. In this case, since winning games depends on the outcomes, we would like to have a way to describe or even predict the likelihood of these outcomes and hence the likelihood of winning. In fact, it is precisely this type of endeavor that sparked the original interest in probability theory. People wanted to know how to gamble and how to win.

Our journey into the theory and practice of probability will begin with discussing the basic objects used to describe experiments and their outcomes. As it turns out, you are already familiar with the mathematical objects that are perfect for their description. These object are called sets (and their members/elements) and you have used them (either explicitly or implicitly) in every math course you've ever taken. For example, you are likely familiar with the set of points in the plane equidistant from the origin. Of course, this is also known as a circle and might be written as

$$\left\{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} = r\right\}$$

where r is some fixed positive number which we call the radius. Another example is the set of natural numbers,

$$\mathbb{N} = \{0, 1, 2, 3, \dots\}.$$

For an example with a clearer probabilistic interpretation, we could consider the set of outcomes of an experiment where a coin is flipped. If the coin is standard, these outcomes are "heads" (which we abbreviate as H) and "tails" (abbreviated T) and so the set of outcomes is $\{H, T\}$. Let's now introduce some standard language and terminology which we will use throughout the course.

Definition 1.1. *Given an experiment with which we can describe distinct outcomes, we shall call the set of all outcomes as the* ***sample space****. The sample space will usually be denoted by Ω and its elements/members' , called* ***outcomes****, are denoted by ω.*

Remark 1.2. If you haven't seen these symbols, both ω and Ω are representations of the Greek letter "Omega". The smaller one, ω, is lower-case Omega and the larger one, Ω, is upper-case Omega.

Remark 1.3. We shall use the standard mathematical symbolism

$$\omega \in \Omega$$

meaning that the outcome ω is a member of (or belongs to) the sample space Ω. This is equivalent to writing $n \in \mathbb{N}$ to symbolize the statement "n is a natural number".

For a given experiment, we should think of the sample space as simply a way of modeling that experiment and its outcomes. The precise way in which we label everything actually shouldn't matter much, of course. This is elucidated by the following example.

Example 1.1: Two Coin Flips

Consider an experiment where we flip two standard coins and keep track of their (individual) outcomes. We could easily model this experiment and its outcomes by writing its sample space as

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$$

where each outcomes $\omega = (\cdot,\cdot)$ is an ordered pair where the first entry denotes the result of the first coin flip and the second denotes the result of the second coin flip. For example, $\omega = (H,T)$ is the outcome that the first coin landed with "heads" up and the second with "tails" up. A clearly equivalent way of writing these events by by juxtaposition, i.e., writing HT for (H,T), which we will often do. In this presentation, we have

$$\Omega = \{HH, HT, TH, TT\}.$$

Another way to describe the same experiment is by assigning the sample space to be

$$\Omega = \{(1,1),(1,0),(0,1),(0,0)\}$$

where, again, the outcomes are given by ordered pairs but where the values of 1 and 0 correspond to heads and tails, respectively. Can you think of other ways to represent a sample space for this experiment?

Example 1.2: N Coin Flips

Generalizing the preceding example, consider the situation in which we flip N standard coins and keep track of their (individual) outcomes. We can represent the sample space for this experiment by

$$\Omega_N = \{(x_1,x_2,x_3,\ldots,x_N) : x_k = H \text{ or } T \text{ for } k = 1,2,\ldots,N\}.$$

In other words, the outcomes of Ω are N-tuples where each entry (coordinate) is simply a H or a T. For example $\omega = (H,T,HH)$ is an outcome of Ω_4, the sample space of 4 coin flips, where "heads" was observed for the first, third and forth flips and "tails" was observed on the second flip. Can you think of a ways to represent this experiment by binary strings?

Example 1.3: Until Heads Appears

Consider the experiment of repeatedly flipping a coin and waiting until the first appearance of "heads". For example, "heads" could appear on the first flip or we could have a sequence of ten flips of "tails" and then "heads" on the eleventh flip. A moment's thought shows that the relevant sample space could be described by binary sequences of the form

$$\Omega = \{H, TH, TTH, TTTH, TTTTH, \ldots\}.$$

In principle, any (integer) number of "tails" could appear before the first "heads" and so we see that this sample space must be infinite. If we were to pay attention to only the number of flips until heads appears, we could write this sample space equivalently by

$$\Omega = \{1,2,3,\ldots\} = \{n \in \mathbb{N} : n > 0\} = \mathbb{N}_+.$$

This is the first (of many) experiments we shall see with infinitely many outcomes.

Given a sample space Ω modeling an experiment, we will often be concerned not just with its outcomes but of certain subsets of its outcomes which we call events. Precisely, we say that E is an **event** if every element ω of E is also an element/outcome of Ω. In this case we write $E \subseteq \Omega$. We can easily see that events allow us to talk about

certain aspects of experiments that may be less fine/precise than talking about individual outcomes. For example, if we are to model the experiment of flipping two coins by

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\},$$

it is natural to talk about the event that the first coin flipped was "heads" and describe this by

$$E = \{(H,H),(H,T)\}.$$

This is, of course, the collection of the outcomes of two flips of a coin where the first outcome is H.

Example 1.4: Rolling Dice

Consider an experiment where we roll two dice and keep track of their (individual) outcomes. We can model this experiment by the sample space

$$\Omega = \{(j,k) : j,k \in \mathbb{N},\ 1 \leq, j,k \leq 6\}.$$

In this model, the values of the first and second die are encoded, respectively, by values j and k running through the natural numbers 1 through 6. Thus, for example, rolling "snake eyes" is precisely the outcome $\omega = (1,1)$.

Perhaps, for the purpose of a game, we are interested in two events E, the event that a 1 is rolled, and F, the event that the sum of faces is exactly 8. Going though all of our possible outcomes $\omega \in \Omega$, we find that

$$E = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$$

and

$$F = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}.$$

Exercise 1.1: The Gambler's Dispute of de Méré

You may not be surprised to learn that the modern theory of probability began as the result of a "Gambler's Dispute". In fact, most of the early drive to understand probability was done to be better at gambling. This initial dispute happened in 1654 when Chevalier de Méré asked the mathematician Blaise Pascal to settle a question concerning a game of chance. In attempting to answer the question, Pascal wrote to the mathematician Pierre de Fermat and their correspondence represents the first known mathematical treatment of a probability theory. In this exercise, we discuss Méré's essential problem.

Suppose that two players, Player A and Player B, are tossing a coin. Each game is simple: They toss a coin and if the coin lands on "heads", Player A wins a point and if the coin is "tails", Player B wins a point. The players have agreed to play this game until the first player has 6 points; at that time, the player with 6 points wins a pot of money. The game, however, unexpectedly stops (and cannot continue) after Player A has won 5 points and player B has won 3 points. Chevalier de Méré asked: Given that the players were not able to finish, how should the pot of money be distributed to accurately reflect the status of the game at the time it was stopped?
Though we won't answer de Méré's question now (we will later), let's think a little about the game.

1. How many possible remaining rounds of the game are there? That is, what is the maximum number of games can be played before there is a winner?
2. Please enumerate the possible outcomes of the game. In other words, write down the sample space.
3. Though we haven't yet talked about probability (we will in the next chapter), do you suspect that the

outcomes are equally likely? Please explain.

4. What is the event that Player A wins? What is the event that Player B wins?

Given a sample space Ω, it should be clear that Ω is a subset of itself, i.e., every $\Omega \subseteq \Omega$, and therefore an event. This can be thought of the event that *something* happens – it doesn't matter what. Also, note that the empty set $\varnothing$, which is the set containing no members does has the property that all of its elements (of which there are none) belong to Ω – this is said to be true by *vacuity*. For this reason, we shall refer to $\varnothing$ as the **empty event** or the **null event**. You can sometimes think of it as representing the experiment not having been performed, i.e., the game not (or not yet) played.

1.1 Operations On Events

Let Ω be a sample space and E and F be events. We say that E is a **subevent (or subset)** of F when $E \subseteq F$, i.e., when every outcome in E is also an outcome in F. If $E \subseteq F$ and $F \subseteq E$, E and F share exactly the same outcomes and so it makes sense to call them equal and write $E = F$. The **union** of the events E and F is defined to be the set of outcomes belonging to E or F inclusively. This is the event

$$E \cup F = \{\omega \in \Omega : \omega \in E \text{ or } \omega \in F\}.$$

More generally, if we have a collection of events $E_1, E_2, E_3, \ldots$, the **union** of these events is the collection of outcomes that belongs to at least one event in the list. This is the event

$$\bigcup_n E_n := \{\omega \in \Omega : \omega \in E_k \text{ for some } k\}.$$

In the case that this list of events is finite, i.e., we are considering the union of $E_1, E_2, \ldots, E_N$ for some N, the above union will also be written $\bigcup_{n=1}^N E_n$. In the case that this list it infinite, we will commonly write the above union as $\bigcup_{n=1}^\infty E_n$. We note that $E \subseteq E \cup F$, $F \subseteq E \cup F$, and, for each k,

$$E_k \subseteq \bigcup_n E_n.$$

Let's now discuss intersections. Given events E and F of a sample space Ω, the **intersection of E and F** is the set of outcomes of Ω that belong to both E and F. This is the event

$$E \cap F = \{\omega \in \Omega : \omega \in E \text{ and } \omega \in F\}.$$

Though not a common notation throughout all of mathematics, in probability and its applications it is customary to write

$$EF = E \cap F.$$

This notation turns out to be pretty handy, as we will see. For a collection of events $E_1, E_2, \ldots$, their common **intersection** is the collection of outcomes that belongs to every E_k in the list. This is the set

$$\bigcap_n E_n = \{\omega \in \Omega : \omega \in E_k \text{ for every } k\}.$$

When this list is finite, i.e., we are considering $E_1, E_2, \ldots, E_N$ for some fixed N, we will simply write this intersection as

$$E_1 E_2 \cdots E_N = \bigcap_n E_N.$$

The following proposition takes care of some basic facts about unions and intersections of events (and it justifies some of our notation above).

Proposition 1.4. *Let E, F, and G be events. We have*

1. $$E \cup F = F \cup E \qquad \text{and} \qquad EF = FE$$
2. $$(EF)G = E(FG) \qquad \text{or, equivalently,} \qquad (E \cap F) \cap G = E \cap (F \cap G)$$
3. $$(E \cup F)G = EG \cup FG$$
4. $$(EF) \cup G = (E \cup G)(F \cup G)$$

We remark that the second assertion above allows us to write EFG or $E \cap F \cap G$ for the common intersection – this is precisely the statement that the intersection is "associative". Similarly, the union of three (or more) sets is also an associative operation. In thinking about the definitions for a moment, it is clear (I hope), that the first and second assertions are valid. Here, I will give two proofs of the third assertion and invite you to think about a proof of the fourth.

Proof #1. I will show that $(E \cup F)G \subseteq EG \cup FG$ and $EG \cup FG \subseteq (E \cup F)G$ and, with both of these containments, the equality must hold. To see the first containment, we choose an arbitrary outcome $\omega \in (E \cup F)G$. This means that ω belongs to the union $E \cup F$ and it also belongs to the event G. Since $\omega \in E \cup F$, $\omega \in E$ or $\omega \in F$ – wherever it is, however, it must belong to G. If ω belongs to E, then it must belong to EG (because it always is in G). Otherwise, it must belong to F and so to FG. Hence ω belongs to EG or it belongs to FG. Hence $\omega \in EG \cup FG$. Since we have shown that an arbitrary outcome ω in $(E \cup F)G$ must also belong to $EG \cup FG$, then it must be the case that $(E \cup F)G \subseteq EG \cup FG$.

It remains to show that $EG \cup FG \subseteq (E \cup F)G$. To see this, let us select some outcome $\omega \in EG \cup FG$. By definition (of the union), $\omega \in EG$ or $\omega \in FG$. It $\omega \in EG$, then $\omega \in E$ and $\omega \in G$. Otherwise, $\omega \in FG$ and so $\omega \in F$ and $\omega \in G$. We note that, in both cases, we have that $\omega \in G$. Taking both cases together, we see that ω must belong to E or F, but it always belongs to G. Hence $\omega \in E \cup F$ and $\omega \in G$. Hence $\omega \in (E \cup F)G$. Just as above, this shows that $EG \cup FG \subseteq (E \cup F)G$ which is what we needed to show. □

Proof #2: This was done in class. □

Exercise 1.2:

Let E, F, and G be events in a sample space Ω.

1. Show that, in general,
$$(E \cup F) \cap G \neq E \cup (F \cap G).$$
In other words, find an example of three events E, F, and G in a sample space Ω (of your choosing) for which $(E \cup F) \cap G$ and $E \cup (F \cap G)$ are not equal.
2. Can you find some general condition on E, F, and G (i.e., a relationship between them) so that
$$(E \cup F) \cap G \neq E \cup (F \cap G)?$$
3. Show that $E \subseteq F$ if and only if $EF = E$.
4. Show that $E \subseteq F$ if and only if $E \cup F = F$.

For an event E in a sample space Ω, **the complement of** E is the set

$$E^c = \{\omega \in \Omega : \omega \notin E\}.$$

For two events, E and F, the **difference between F and E** is the set

$$F \setminus E = \{\omega \in F : \omega \notin E\}.$$

For a couple of nice characterizations of set differences, we have the following proposition.

Proposition 1.5. *For events E and F, we have the identities*

$$E^c = \Omega \setminus E \qquad \text{and} \qquad F \setminus E = FE^c.$$

Proof. In looking at the definitions of E^c and $\Omega \setminus E$, it is not hard to see that they are identical – so there is nothing to prove here. For the other identity, we will show that

$$F \setminus E \subseteq FE^c \qquad \text{and} \qquad FE^c \subseteq F \setminus E$$

from which we get equality. To see the first so-called containment, let's take an event $\omega \in F \setminus E$. In this case, $\omega \in F$ and $\omega \notin E$. Since ω must be a member of Ω (given that it's an outcomes) and $\omega \notin E$, we have $\omega \in E^c$. Thus $\omega \in F$ and $\omega \in E^c$ and so it belongs to the intersection $F \cap E^c = FE^c$. Hence $F \setminus E \subseteq FE^c$.

To see the reverse containment, suppose the $\omega \in FE^c$. In this case, $\omega \in F$ and $\omega \in E^c$. Since ω belongs to E^c, it cannot belong to E, i.e., $\omega \notin E$. Thus, $\omega \in F$ but $\omega \notin E$ and so, by definition, $\omega \in F \setminus E$. We have thus shown that $FE^c \subseteq F \setminus E$ and, together with the previous containment, we have shown the events are equal. □

The following proposition, known as the famous "De Morgan's Laws" illustrates the interplay between unions and intersections under complements.

Proposition 1.6. *Let E and F be events in a sample space Ω. Then*

$$(E \cup F)^c = E^c \cap F^c$$

and

$$(E \cap F)^c = E^c \cup F^c.$$

More generally, if $E_1, E_2, \ldots$ is a collection of events in Ω, then

$$\left(\bigcup_n E_n\right)^c = \bigcap_n E_n^c$$

and

$$\left(\bigcap_n E_n\right)^c = \bigcup_n E_n^c.$$

Here I'll prove the two-event version of De Morgan's laws above, i.e., that for E and F. To test your understanding, I invite you to try the proof for three events or a more general collection, $E_1, E_2, \ldots,$.

Proof. Let's first show that $(E \cup F)^c = E^c \cap F^c$ by showing that $(E \cup F)^c \subseteq E^c \cap F^c$ and the reverse containtment, $F^c \cap E^c \subseteq (E \cup F)^c$. To see the first containment, let ω be an element of $(E \cup F)^c$. By definition of the complement, ω cannot belong to $E \cup F$ and so ω cannot be a member of E or F otherwise it would belong to their union. Hence ω must live in both E^c and F^c and so ω belongs to $E^c \cap F^c$. We have thus showed that $(E \cup F)^c \subseteq E^c \cap F^c$.

To see the reverse containment, let's take an arbitrary element ω of $E^c \cap F^c$. Then ω must belong to both E^c and F^c. Thus ω cannot belong to E and ω cannot belong to F, by definition of the complement. Hence, ω cannot belong to the union $E \cup F$ for, otherwise, it could belong to at least one of these events. Thus ω must be a member of $(E \cup F)^c$. We have thus shown that $E^c \cap F^c \subseteq (E \cup F)^c$.

Let's now show the other De Morgan law for E and F. We could argue this directly like we did for the first De Morgan law, but there is an easier way to do it. Let's apply the first De Morgan law to E^c and F^c. This gives the identity of events,

$$(E^c \cup F^c)^c = (E^c)^c \cap (F^c)^c.$$

Of course, the complement of the complement is the event itself, so $(E^c)^c = E$ and $(F^c)^c = F$. Hence, the above identity becomes

$$E \cap F = (E^c \cup F^c)^c.$$

If we take the complement of both sides (which also have to be equal), we have

$$(E \cap F)^c = ((E^c \cup F^c)^c)^c = E^c \cup F^c$$

which is precisely the second De Morgan law. □

Given two events E and F, we say that E **and** F **are disjoint** if they share no common elements. In other words, E and F are disjoint if $EF = \varnothing$. A collection of events $E_1, E_2, \ldots$ is said to be **pairwise disjoint** (or the sets are said to be pairwise disjoint) provided that $E_j E_k = \varnothing$ for every $j \neq k$. Sometimes we shall simply say that these sets are disjoint. Using the notion of pairwise disjoint collections, we shall find the following definition of "partition" extremely useful.

Definition 1.7. *A collection of events $E_1, E_2, \ldots$ in a sample space Ω is said to be a **partition of** Ω if the events are pairwise disjoint and their union is all of Ω, i.e., if they satisfy the following two conditions:*

1. $E_j E_k = \varnothing$ *whenever* $j \neq k$.
2. $\Omega = \bigcup_n E_n$.

Example 1.5: Partitions of Two Coin Flips

Consider an experiment in which we flip two coins. As we've previously discussed, an appropriate sample space for this experiment is

$$\Omega = \{HH, HT, TH, TT\}.$$

There are several partitions we can form for this sample space. Consider, for example,

$$E_1 = \{HH\}, \quad E_2 = \{HT\}, \quad E_3 = \{TH\}, \quad \text{and} \quad E_4 = \{TT\}.$$

It is clear by construction that no events E_j and E_k share a common element and so this collection is pairwise disjoint. Also, given that each member of Ω belongs to one (and in this case only one) event E_k, we have $\Omega = E_1 \cup E_2 \cup E_3 \cup E_4$ and so this collection forms a partition.
The above is, of course, not the only partition. Consider instead the events

$$F_1 = \{HH, HT, TH\} \qquad \text{and} \qquad F_2 = \{TT\}.$$

It is easy to check the two requisite conditions to see that they forms a partition of Ω.

Exercise 1.3: Indicators

Given an event E in a sample space Ω, the **indicator function of** E is the function $\mathbb{1}_E : \Omega \to \mathbb{R}$

$$\mathbb{1}_E(\omega) = \begin{cases} 1 & \omega \in E \\ 0 & \omega \notin E \end{cases}$$

As we will see, the indicator function is useful in understanding the events E and, later, their probabilities. In this exercise, we see that they are useful in understanding operations on events.

1. Given events E and F, show that $\mathbb{1}_E \mathbb{1}_F = \mathbb{1}_{EF}$. In other words, show that

$$\mathbb{1}_E(\omega)\mathbb{1}_F(\omega) = \mathbb{1}_{EF}(\omega).$$

for all $\omega \in \Omega$.

2. Express $\mathbb{1}_{F\setminus E}$ in terms of $\mathbb{1}_E$ and $\mathbb{1}_F$.
3. Show that $\mathbb{1}_{E\cup F} \leq \mathbb{1}_E + \mathbb{1}_F$.
4. Under what conditions (on E and F) is it true that $\mathbb{1}_{E\cup F} = \mathbb{1}_E + \mathbb{1}_F$? Justify it.

Note here.

1.2 A note on the size of sets

There are many different kinds of experiments we will discuss throughout this course. As we have already seen, we can consider experiments with a finite number of outcomes. For example, we've already talked (a lot) about the two-element sample space $\Omega = \{H, T\}$ representing the flip of a single coin. Another such experiment with a finite number of outcomes is the experiment in which we select three cards from a deck (of fifty two) at random without replacement. As we will see in Chapter 3, this experiment has a sample space Ω with $52 \times 51 \times 50 = 132600$ outcomes! We have also considered experiments with an infinite number of outcomes. For examples, we discussed the example in which a coin was flipped repeatedly until the first appearance of "heads" and there we say a sample space $\Omega = \mathbb{N}_+$ with infinitely many outcomes. Consider now the game of (continuous) plinko:

Example 1.6: Continuous Plinko

Suppose that we play a game of plinko in which a marble is dropped in a box of horizontal dimension 1 meter. In the process of falling to the bottom of the box, it hits "pegs" on the way down and eventually lands at the bottom of the box where any position $0 \leq x \leq 1$ is, in principle, possible. This differs from the classic game of PLINKO where there are only finitely many possible locations (or bins), i.e., we're playing plinko with the bin walls removed. Note.

We can model continuous plinko by the sample space as the unit interval, i.e.,

$$\Omega = \{x \in \mathbb{R} : 0 \leq x \leq 1\} = [0, 1].$$

Note that this interval contains every rational number of the form $1/n$ for $n \in \mathbb{N}_+$, and so it must be infinite. In fact, it's not difficult to see that every rational number between 0 and 1 is also contained in Ω, i.e.,

$$\left\{\frac{n}{m} : n \in \mathbb{N}, m \in \mathbb{N}_+, n \leq m\right\} = \mathbb{Q} \cap [0, 1] \subseteq \Omega.$$

But, you should note that not every number in Ω is rational. For example, $\sqrt{2}/2$ and $\pi/4$ are irrational numbers in Ω. What's, perhaps, not clear is that Ω contains (way!) more irrational numbers than it does rational ones. This leads to the idea the the infinite nature of $\Omega = [0, 1]$ differs from the nature of the infinity that is $\mathbb{N}_+$.

In light of the above examples, we will find it useful for us to talk about the size of certain sample spaces and events in terms of the number of outcomes they contain. For this, we have the following definition.

Definition 1.8. *Let A be a set (or event).*

1. *We say that that a set A is **finite** if it contains a finite number of elements. In other words, A is finite if the number of elements it contains (denoted by $\#(A)$) is some number $N \in \mathbb{N}$. In this case, we may write*

$$A = \{a_1, a_2, \ldots, a_N\}.$$

2. *We say that A is **countably infinite** if it is in one-to-one correspondence with the natural numbers, $\mathbb{N}$. In other words, there is a bijection (a one-to-one and onto function) $f : \mathbb{N} \to A$ and, the existence of such a function allows us to label A as*
$$A = \{a_1, a_2, a_3, \dots\}$$
where the above list doesn't end and every member of A is of the form a_n for some n. In this case, we shall write $\#(A) = \aleph_0$; here, $\aleph$ is the first letter of the Hebrew alphabet and is pronounced "aleph" and we pronounce $\aleph_0$ as "aleph-zero".

3. *We say that A is a **countable** set if it is finite or countably infinite. If A is not countable, we say that A is **uncountable**.*

In this course, we shall not work technically with the study of the size of sets (cardinality). Cardinality is often a subject covered in MA274 and, if this topic interests you, I strongly encourage you to take MA274 in the near future (or a course in set theory). The only thing that we will need in this course is the proposition, which we shall take for granted.

Proposition 1.9. *The following statements are true.*

1. *The sets $\mathbb{N}_0, \mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are all countably infinite and hence countable.*

2. *The unit interval $[0, 1]$ (or any interval $[a, b]$ with $a < b$), the real line $\mathbb{R}$, the plane $\mathbb{R}^2$, the unit disk $B = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$, and the unit square $[0, 1] \times [0, 1]$ (or any rectangle of the form $[a, b] \times [c, d]$ with $a < b$ and $c < d$) are uncountable. In fact, all of the above mentioned have the same cardinality (i.e., they are in one-to-one correspondence).*

HERE

Dart Throwing

Dart throwing

Chapter 2

Modeling Uncertainty: Probability Measures

In the previous chapter, we discussed how to describe an experiment and its outcomes. The focus of this chapter is to develop a way in which we can describe the likelihood of outcomes. For example, we may consider the experiment of flipping a coin and, if we are reasonably convinced that the coin is fair (i.e., its faces have the same weight), we might say that the likelihood of getting "heads" and "tails" is the same represent both by the value 1/2. We might call this value the probability of getting "heads" or "tails". Unfortunately, we are immediately led toward a philosophical conundrum: How can we possibly know that the coin is fair (or the toss is fair)? In the practice of the (single) experiment, either the coin lands on "heads" or it lands on "tails" and so there is no way to know whether these outcomes were equally likely. We could perhaps flip the coin over and over again and in a way where each flip is made independently[1] of those before it and we could count the number of heads versus the number of total flips. Precisely, we could flip the coin n times and consider the ratio

$$Q_n(H) = \frac{\text{Number of "heads" in } n \text{ flips}}{n}.$$

If the coin is fair, we would certainly expect that this ratio approach $1/2$ as $n \to \infty$. But, of course, our lives finite and we can't flip forever. When do we stop? How exactly supposed to investigate if $\lim_{n\to\infty} Q_n(H)$ is $1/2$ or if the limit exists at all?

What we've introduced above is the so-called frequentist definition of probability and, though it is intuitive, it is fraught with the type of issues discussed above. To get around this issue and to introduce an interpretation that is meaningful (and testable), we shall simply model the probability of the experiment we hope to describe. In other words, we shall write down a theoretical description of probability and work out the theory's implications and consequences; this is the nature of mathematics. It is the job of statistics (or experiment) to test and analyze whether or not our model is reasonable. What we're describing here is precisely the difference between theory and experiment and, in this course, we shall be concerned with the former.

2.1 Probability Measures

In this section, we make our first attempt at modeling an experiment by introducing the notion of probability measure. As we shall see, our definition below is intuitive and it works perfectly for experiments with a finite

[1] Here, we are using the term "independent" to mean that each flip is done in a way which is completely insensitive to the flips before it. This might agree with your understanding of the word independent in common English – I hope that it is. In the next chapter, we will present a rigorous definition of the term "independent" and show that it agrees with our current understanding. We will also see ways in which the definition is somewhat unintuitive.

number of outcomes. In general, the definition we give below will need to be refined so that it makes sense in the case of infinite sample spaces. This refinement will be given in time. Our first attempt is as follows.

Definition 2.1 (Probability Measure – A first attempt)**.** *Let Ω be a non-empty sample space. A probability measure $\mathbb{P}$ is an assignment of the events of Ω to real numbers which satisfies the following three conditions:*

1. $\mathbb{P}(E) \geq 0$ *for every event E.*
2. *For any two disjoint events E and F, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$.*
3. $\mathbb{P}(\Omega) = 1$.

We shall sometimes refer to these conditions as the "axioms" of probability. The second axiom is sometimes called ***pairwise additivity****.*

Example 2.1: Coin Flips

Consider the experiment of flipping a single coin with sample space $\Omega = \{H, T\}$. Let's define $\mathbb{P}$ to be the function

$$\mathbb{P}(E) = \begin{cases} 0 & E = \varnothing \\ \frac{1}{2} & E = \{H\}, \{T\} \\ 1 & E = \Omega \end{cases}$$

for each event $E \subseteq \Omega$. We remark that $\mathbb{P}$ assigns equal probabilities to the single-outcome events $\{H\}$ and $\{T\}$, consistent with the interpretation that our coin is fair. To verify that $\mathbb{P}$ is a bona fide probability measure, it remains to check that conditions 1-3 of the definition are satisfied by $\mathbb{P}$. Since $\mathbb{P}$ only takes the values 0, 1/2, and 1 and $\mathbb{P}(\Omega) = 1$, it is easy to see that $\mathbb{P}$ satisfied both the first and third axiom. To see the second axiom, we first remark that there are only a handful of events that can be chosen which are disjoint. First, if E is any of the events, $\varnothing$, $\{H\}$, $\{T\}$, or Ω, and $F = \varnothing$, it is clear that E and F are disjoint and that $E \cup F = E$. So,

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) = \mathbb{P}(E) + 0 = \mathbb{P}(E) + \mathbb{P}(F)$$

as required. If we do not involve the nullevent, there are only two disjoint events left at $E = \{H\}$ and $F = \{T\}$. In this case,

$$\mathbb{P}(E \cup F) = \mathbb{P}(\Omega) = 1 = \frac{1}{2} + \frac{1}{2} = \mathbb{P}(E) + \mathbb{P}(F)$$

and so we have verified the second axiom of probability measure. Hence, $\mathbb{P}$ is a probability measure.

The probability measure defined above is not the only one we can put on the coin-flip sample space Ω. Consider alternatively the function $\mathbb{P}^*$ defined by

$$\mathbb{P}^*(E) = \begin{cases} 0 & E = \varnothing \\ \frac{2}{3} & E = \{H\} \\ \frac{1}{3} & E = \{T\} \\ 1 & E = \Omega \end{cases}$$

for each event $E \subseteq \Omega$. By precisely the same argument given above, it is clear that $\mathbb{P}^*$ is also probability measure. This is the measure which models the situation in which "heads" is twice as likely as "tails".

As discussed above, $\mathbb{P}$ in the above example assigns all singleton-outcome events as equally likely. As the following proposition shows, every finite sample space has such a probability measure.

Proposition 2.2. *Let Ω be a non-empty but finite sample space and define*

$$\mathbb{P}_u(E) = \frac{\#(E)}{\#(\Omega)}$$

for each event $E \subseteq \Omega$. Then $\mathbb{P}_u$ is a probability measure on Ω. Further, it assigns the same probability to all single-outcome events, i.e., for any two outcomes $\omega_1, \omega_2 \in \Omega$,

$$\mathbb{P}_u(\{\omega_1\}) = \frac{1}{\#(\Omega)} = \mathbb{P}_u(\{\omega_2\})$$

*We shall refer to this measure as the **equally likely** measure or the **uniform probability measure** on Ω.*

Proof. First, let's define $\mathbb{P}_u(E) = \#(E)/\#(\Omega)$ where, as usual, $\#(E)$ denotes the number of outcomes in E. By this definition, it is clear that $\mathbb{P}_u$ is non-negative and has $\mathbb{P}_u(\Omega) = 1$. Thus, to verify that $\mathbb{P}$ is a probability measure, we must verify the second axiom of Definition 2.1. To this end, let E and F be disjoint events in Ω. In the case that either event is the nullevent, verifying pairwise additivity can be done precisely as we did in the preceding example. Let's therefore assume that E and F are both nonempty (and still disjoint). We may label their elements as

$$E = \{\omega_1, \omega_2, \ldots, \omega_M\} \qquad \text{and} \qquad F = \{\eta_1, \eta_2, \ldots, \eta_N\}$$

where $\#(E) = M$ and $\#(F) = N$. Given that E and F share no common elements, non of the ωs or ηs are the same and so

$$E \cup F = \{\omega_1, \omega_2, \ldots, \omega_M, \eta_1, \eta_2, \ldots, \eta_N\}.$$

In particular, $E \cup F$ has $M + N$ elements and so

$$\mathbb{P}_u(E \cup F) = \frac{\#(E \cup F)}{\#(\Omega)} = \frac{M+N}{\#(\Omega)} = \frac{M}{\#(\Omega)} + \frac{N}{\#(\Omega)} = \mathbb{P}_u(E) + \mathbb{P}_u(F).$$

Thus, $\mathbb{P}_u$ is a probability measure. Finally, observe that, for each outcome ω, we have

$$\mathbb{P}_u(\{\omega\}) = \frac{1}{\#(\Omega)}$$

and so $\mathbb{P}_u$ assigns the same probability to every single-event outcome. □

Exercise 2.1: From Pairwise Additivity to Finite Additivity

Let Ω be a non-empty but finite sample space and $\mathbb{P}$ be a probability measure

1. Show that, given any finite (pairwise) disjoint collection of events $E_1, E_2, \ldots, E_M$,

   $$\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_M) = P(E_1) + \mathbb{P}(E_2) + \cdots + \mathbb{P}(E_M).$$

   We shall call this property **countable additivity**. Hint: You can argue by simply applying pairwise additivity over and over again. If you are comfortable doing mathematical induction, do that.

2. Consider now the uniform measure $\mathbb{P}_u$ on Ω given by $\mathbb{P}_u(E) = \#(E)/\#(\Omega)$. In the above proposition, we showed that $\mathbb{P}_u$ assigns the same probability to all single event outcomes. Here, you show that $\mathbb{P}_u$ is the only such measure, i.e., that the uniform measure is unique. To this end, assume that $\mathbb{P}$ is a probability measure and that there is a fixed constant $p > 0$ for which

   $$\mathbb{P}(\{\omega\}) = p$$

   for all $\omega \in \Omega$. Use countable additivity to show that, in fact, $\mathbb{P} = \mathbb{P}_u$. Hint: If $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ with $\#(\Omega) = N$, then Ω can be partitioned by the events $\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_N\}$. Use this to deduce that $p = 1/N$ and, with this, argue that $\mathbb{P}$ and $\mathbb{P}_u$ must agree.

Example 2.2: Two (and N) coin flips

Consider the sample space

$$\Omega = \Omega_2 = \{HH, HT, TH, TT\}$$

corresponding to two coin flips (or flipping the same coin twice and keeping track of the result). Observe that the uniform probability measure gives

$$\mathbb{P}_u(\{HH\}) = \mathbb{P}_u(\{HT\}) = \mathbb{P}_u(\{TH\}) = \mathbb{P}_u(\{TT\}) = \frac{1}{4} = \frac{1}{2^2}.$$

More generally, consider the sample space of N coin flips,

$$\Omega_N = \{(x_1, x_2, \ldots, x_N) : x_k = H \text{ or } T \text{ for all } k = 1, 2, \ldots, N\}.$$

As we will see in the next chapter on counting, $\#(\Omega_N) = 2^N$ and therefore the uniform probability measure on Ω_N has

$$\mathbb{P}_u(\{\omega\}) = \frac{1}{2^N}$$

to each $\omega \in \Omega_N$. As we shall soon argue, the uniform measure correspond to the situation in which all of the coins are flipped independently or, equivalently, we flip a single fair coin N times but each subsequent flip is done independently from those which proceeded it.

As the following result shows, probability measures on finite sample spaces can be defined, equivalently, by simply specifying the measure of each single-outcome event. It also gives a full characterization of probability measures on such sample spaces.

Proposition 2.3. *Let Ω be a non-empty but finite sample space and, for simplicity, let's enumerate its outcomes so that*

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$$

where $N = \#(\Omega)$. Given any collection of non-negative numbers $p_1, p_2, \ldots, p_N$ such that

$$\sum_{k=1}^{N} = p_1 + p_2 + \cdots + p_N = 1,$$

there is a unique probability measure $\mathbb{P}$ on Ω with

$$\mathbb{P}(\{\omega_k\}) = p_k$$

for $k = 1, 2, \ldots, N$.

Proof. Let's define

$$\mathbb{P}(E) := \sum_{\{k:\omega_k \in E\}} p_k.$$

for each event $E \subseteq \Omega$. This summation notation means that we include p_k in our sum is $\omega_k \in E$. Let's show that $\mathbb{P}$ is indeed a probability measure. First, it is easy to see that $\mathbb{P}$ is non-negative since we have required the numbers $p_1, p_2, \ldots, p_N$ to be non-negative. To see that $\mathbb{P}$ satisfies the second axiom, let E and F be events for which $EF = \varnothing$. In this case, we have

$$\mathbb{P}(E) = \sum_{\{k:\omega_k E\}} p_k \qquad \text{and} \qquad \mathbb{P}(F) = \sum_{\{k:\omega_k F\}} p_k.$$

Now, since E and F are disjoint, the first summation is done over a distinct collection of ks than the second is – this is precisely because each ω_k can belong to at most one event E or F. Consequently, if we add these sums we

get precisely the sum of those k which are in E or in F. In other words,

$$\mathbb{P}(E)+\mathbb{P}(F)=\sum_{\{k:\omega_k\in E\}}p_k+\sum_{\{k:\omega_k\in F\}}p_k=\sum_{\{k:\omega_k\in E\cup F\}}p_k=\mathbb{P}(E\cup F)$$

and so we have satisfied the second axiom. Finally, we have

$$\mathbb{P}(\Omega)=\sum_{\{k:\omega_k\in\Omega\}}p_k=\sum_{k=1}^{N}p_k=1$$

where we have used the condition that the total sum of p_ks is 1. Thus, $\mathbb{P}$ is a probability measure. Furthermore, by its definition we see that

$$\mathbb{P}(\{\omega_k\})=p_k$$

for each $k=1,2,\ldots,N$. It remains to show that it is the only measure which satisfies this condition, i.e., that is is the unique probability measure satisfying this condition. Let $\widetilde{\mathbb{P}}$ be a (possibibly) different probability measure having

$$\widetilde{\mathbb{P}}(\{\omega_k\})=p_k$$

for each $k=1,2,\ldots,N$ and let E be an event. In the case that E is non-empty, we can partition the event E into single-outcome events,

$$\mathbb{E}=\{\omega_{k_1}\}\cup\{\omega_{k_2}\}\cup\cdots\cup\{\omega_{k_M}\}=\bigcup_{\{k:\omega_k\in E\}}\{\omega_k\}$$

where $M=\#(E)$. By the finite additivity of $\widetilde{\mathbb{P}}$, we have

$$\widetilde{\mathbb{P}}(E)=\sum_{j=1}^{M}\widetilde{\mathbb{P}}(\{\omega_{k_j}\})=\sum_{\{k:\omega_k\in E\}}\widetilde{\mathbb{P}}(\{\omega_k\})=\sum_{\{k:\omega_k\in E\}}p_k$$

but, of course, this is precisely the definition of $\mathbb{P}(E)$. Thus $\widetilde{\mathbb{P}}(E)=\mathbb{P}(E)$ for every event $E\subseteq\Omega$ and so $\widetilde{\mathbb{P}}=\mathbb{P}$. □

The above proposition shows that each probability measure on a finite sample space is determined uniquely by the probabilities of individual outcomes. One can see that the uniform measure arises exactly when all of these outcomes have the same probability (when all p_k have the value $1/\#(\Omega)$). Going forward, when working on finite sample spaces (or countable ones, as we shall see), we may introduce a probability measure by simply specifying its value on single-outcomes events instead giving a definition of $\mathbb{P}(E)$ for every event E. For example, looking back to Example 2.1, notice that the first uniform/fair-coin measure could have been described by simply specifying that $\mathbb{P}(\{H\})=\mathbb{P}(\{T\})=1/2$. The unfair coin measure could have been described by specifying that $\mathbb{P}(\{H\})=2/3$ and $\mathbb{P}(\{T\})=1/3$.

Exercise 2.2: The Gambler's Dispute of de Méré

Two players, A and B, are tossing a coin. Each game is simple: They toss a coin and, if the coin lands on "heads", Player A wins a point and, if instead the coin is "tails", Player B wins a point. The players have agreed to play this game until the first player has 6 points; at that time, the player with 6 points wins a pot of money. The game, however, unexpectedly stops (and cannot continue) after Player A has won 5 points and player B has won 3 points. Chevalier de Méré asked: Given that the players were not able to finish, how should the pot of money be distributed to accurately reflect the status of the game at the time it was stopped?

As you likely found in Exercise 1.1, an appropriate 4-outcome sample space can be written

$$\Omega_0=\{A,BA,BBA,BBB\}.$$

For example, A is the outcome of the game where heads came up on the first toss and so A won. BA

represents the outcome where tails came up on the first toss and then heads came up on the second toss and so player A still won. Of course, as you suspected, the outcomes in this sample space are not equally likely and so, in particular, the probability that player B wins is not $1/4$.

The following steps, which you will do, is a way to correctly determine the probabilities of the game.

1. First, let's represent the game by a different sample space. Suppose that the players decided to toss the coin three times regardless of whether or not the game should stop because the coin landed on heads before the end of the three flips (wherein Player A would win). This new sample space, Ω_1, has 8 outcomes. What is Ω_1 explicitly? Explain.
2. In this game of three tosses, all outcomes/singleton events are equally likely (i.e., with probability $1/8$). Please explain why this is reasonable.
3. By mapping the outcomes in Ω_1 back to those in Ω_0, you should be able to give the "correct" probabilities of the outcomes in Ω_0. Write down a probability measure on Ω_0 by specifying its value on each of its four outcomes.
4. Using your new probability measure, what is the probability that A wins? What is the probability that B wins?
5. If the pot of money contained \$80, how should it be distributed according to de Méré's query?

There are many properties of a general probability measure that we can establish directly from the axioms. The property below is a handy identity which is a special case of the so-called inclusion-exclusion rule.

Proposition 2.4 (Inclusion-Exclusion)**.** *Let $\mathbb{P}$ be a probability measure on a sample space Ω. For any events E and F, we have*

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF).$$

Proof. Observe that

$$E = EF \cup (E \setminus F)$$

since E consists of those elements which are in E and F or are in E but not F. It is clear that the sets EF and $E \setminus F$ are disjoint and so by the pairwise disjointness of $\mathbb{P}$, we have

$$\mathbb{P}(E) = \mathbb{P}(EF \cup (E \setminus F) = \mathbb{P}(EF) + \mathbb{P}(E \setminus F).$$

By a similar argument, we can write $E \cup F = F \cup (E \setminus F)$ where the events F and $E \setminus F$ are disjoint. Therefore

$$\mathbb{P}(E \cup F) = \mathbb{P}(F) + \mathbb{P}(E \setminus F).$$

Combining the two preceding equations, we find that

$$\mathbb{P}(E \cup F) = \mathbb{P}(F) + \mathbb{P}(E \setminus F) = \mathbb{P}(F) + (\mathbb{P}(E) - \mathbb{P}(EF)) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF)$$

which is what we wanted to show. □

Exercise 2.3: Some Properties – Deduction from the Axioms

Let $\mathbb{P}$ be a probability measure on a sample space Ω and let E and F be events.

1. Show that $\mathbb{P}(E) \leq \mathbb{P}(F)$ whenever $E \subseteq F$. Hint: Write F as the disjoint union of E and $F \setminus E$. This property is called **Monotonicity of Measure**
2. Show that $\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F)$. Hint: Use the proposition above.. This property is a special case

of **Boole's Inequality**.

3. Show that $\mathbb{P}(EF) \geq \mathbb{P}(E) + \mathbb{P}(F) - 1$. Hint: Use the proposition above.

As we discussed, it turns our the our first attempt at the definition of probability measure is problematic. Still, throughout the proceeding examples, we have built some (very correct) intuition for probability measures and it is high time that we give (close to) the actual definition. This is as follows.

Definition 2.5. *Probability Measure Let Ω be a non-empty sample space. A probability measure $\mathbb{P}$ is an assignment of the events of Ω to the real numbers which satisfies the following three properties:*

1. *$\mathbb{P}(E) \geq 0$ for every event E.*
2. *Given any countable collection of pariwise disjoint events $E_1, E_2, \ldots,$,*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(E_n) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \cdots.$$

   *This is known as **countable additivity**.*

3. *$\mathbb{P}(\Omega) = 1$.*

*We shall refer to these conditions as the **axioms of probability**. Henceforth, these axioms should supersede those of our initial definition (the first attempt).*

Remark 2.6. As it turns out, there are sample spaces Ω for which one cannot define a probability measure which is defined on *all* subsets of Ω but only those which we call "measurable". Thus, in general, we must make a distinction between subsets and events and require all events to be so-called measurable. This distinction won't be very important to us and we can generally ignore it. If you're interested, the details of the existence of non-measurable sets are taken up in a course in analysis (e.g. MA339).

When the term *probability measure* is used henceforth, we shall always mean the definition above. Still, it should be noted that I haven't misled you. In the case that the sample space Ω is finite, the two definitions agree. In fact, in the case that Ω is countable, there is essentially no change to the theory that we've discussed so far. For example, the following proposition extends Proposition 2.3 to countable sample spaces.

Proposition 2.7. *Let Ω be countably infinite sample space enumerated as*

$$\Omega = \{\omega_1, \omega_2, \ldots\}.$$

Given any collection of non-negative numbers $p_1, p_2, \ldots,$ with

$$\sum_{k=1}^{\infty} p_k = 1,$$

there is a unique probability measure $\mathbb{P}$ on Ω with

$$\mathbb{P}(\{\omega_k\}) = p_k$$

for all $k = 1, 2, \ldots$.

Example 2.3: Until Heads Appears NOTE

Consider the experiment of repeatedly flipping a coin and waiting until the first appearance of "heads". As

we previously discussed, the relevant sample space for this experiment can be written by

$$\Omega = \{1, 2, \dots, \} = \mathbb{N}_+$$

where each integer $k \in \Omega$ represents the number of flips until "heads" appears. In the case that the coin is fair, it is reasonable to expect that

$$\mathbb{P}(\{1\}) = \frac{1}{2},$$

i.e., the probability that the game stops after one flip is exactly $1/2$ because the coin is fair. For $k = 2$, I claim that (provided the flips are independent – at least according to our colloquial understanding of the term),

$$\mathbb{P}(\{2\}) = \frac{1}{4}.$$

To see that this is reasonable, think of the coin being flipped twice (and not worry about it being flipped further). In this case, we have the outcomes $\{HH, HT, TH, TT\}$ which are all equally likely if the coin is fair and flipped independently. Thus, each of these outcomes has probability $1/4$. The first two outcomes HH and HT represent the cases in which we would have stopped on the first flip, and so their collective probability is $1/2$, which is precisely the probability $\mathbb{P}(\{1\})$. In looking at the final outcomes, TH and TT, the first represents the case in which the game stops exactly after the second flip and TT represents the outcome that we go on to the third or more flips. Thus, we identify $\{2\}$ with the event $\{TH\}$ and hence assign its probability to be $1/4$. Pushing this argument further, it is reasonable to assign

$$\mathbb{P}(\{k\}) = \frac{1}{2^k}$$

for each $k = 1, 2, \dots$. When we introduce the mathematical definition of independence, we give another argument for why this assignment is correct. In view of the proceeding proposition, to confirm that this assignment specifies a probability measure, we must simply verify that

$$\sum_{k=1}^{\infty} \mathbb{P}(\{k\}) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

To see this, we simply remark that this is a geometric series and so it converges precisely when its common ratio $1/2 < 1$. This is, indeed the case and so we have the summation formula

$$\sum_{k=1}^{\infty} \frac{1}{2^k} = \sum_{k=1} \left(\frac{1}{2}\right)^k = \frac{1/2}{1 - (1/2)} = 1.$$

Thus, this assignment of $\mathbb{P}$ determines a probability measure on the "until first heads" sample space and does so uniquely.

Let's consider our first example of a probability measure on an uncountable sample space.

Example 2.4: Continuous Plinko

Let's return to the game of continuous Plinko discussed in the first chapter. We had

$$\Omega = [0, 1]$$

where each $0 \leq \omega \leq 1$ represented the landing position of a marble in the box after hitting many pegs on

the way down. On this sample space, let's define the measure

$$\mathbb{P}(E) = \int_0^1 \mathbb{1}_E(x)\,dx$$

for each event[a] E. For example, consider the event that the marble lands in the left half of the box. This is the event

$$E = [0, 1/2]$$

and the graph of the associated associated indicator function $\mathbb{1}_{[0,1/2]}$ is shown in Figure 2.1.

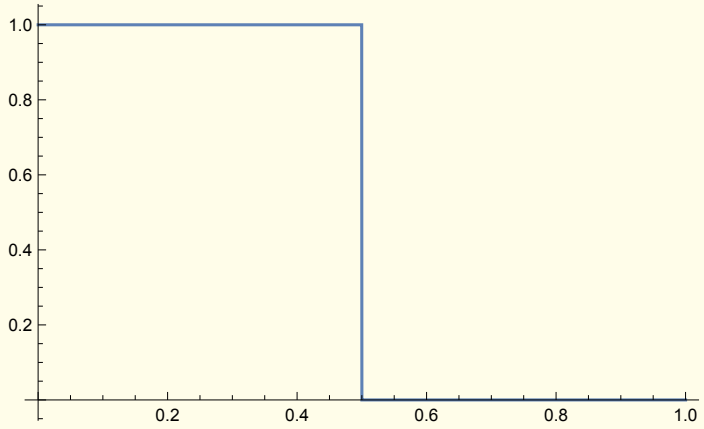


Figure 2.1: Graph of $\mathbb{1}_{[0,1/2]}$.

By taking the integral to be the area under the graph, we see that

$$\mathbb{P}([0, 1/2]) = \int_0^1 \mathbb{1}_{[0,1/2]}(x)\,dx = \int_0^{1/2} 1\,dx = \frac{1}{2}.$$

Taking this idea slightly further, for any event E of the form $E = [a, b]$ (or (a, b), $(a, b]$, $[a, b)$) for $0 \leq a \leq b \leq 1$, we have

$$\mathbb{P}(E) = \int_a^b 1\,dx = b - a = \text{length}(E).$$

With this property, we see that $\mathbb{P}$ assigns equal probability to intervals of equal lengths. For this reason, as we will see later, we will call this a uniform measure on the uncountable sample space Ω.

To see that this does indeed define a probability measure, we need to verify the three conditions. Seeing the first condition is relatively straightforward. Since the indicator function is always non-negative (it's either 1 or 0), we have

$$\mathbb{P}(E) = \int_0^1 \mathbb{1}_E(x)\,dx \geq \int_0^1 0\,dx = 0.$$

Hence $\mathbb{P}(E) \geq 0$ for all events E. For the third condition, we see that

$$\mathbb{P}(\Omega) = \mathbb{P}([0, 1]) = \text{length}([0, 1]) = 1$$

and so the third condition is satisfied. It remains to show the second condition and, as it turns out, checking this condition is fairly difficult (and is really part of the subject of a course in measure theory). Still, I will give an argument here which is not quite correct but I hope you will find convincing:

A not-quite-correct proof of countable additivity. Let $E_1, E_2, \ldots$ be disjoint events (actually, for simplicity, assume that they are all intervals like we've looked at above).I claim that, for $E = \cup_n E_n$,

$$\mathbb{1}_E(x) = \sum_{n=1}^{\infty} \mathbb{1}_{E_n}(x)$$

for all $x \in \Omega$. To verify that this identity, notice that each $x \in \Omega$ is either in $E = \cup_n E_n$ or it isn't. If it is not in E (i.e., then $x \in \Omega \setminus E$), then x cannot be in any E_n for any n and hence

$$\sum_{n=1}^{\infty} \mathbb{1}_{E_n}(x) = \sum_{n=1}^{\infty} 0 = 0 + 0 + \cdots = 0 = \mathbb{1}_E(x).$$

In the case that $x \in E = \cup_n E_n$, then $x \in E_{n'}$ for some n'. Of course, since the events are disjoint, x can only belong to this single event $E_{n'}$. Hence

$$\mathbb{1}_{E_n}(x) = \begin{cases} 1 & n = n' \\ 0 & n \neq n' \end{cases}.$$

Correspondingly,

$$\begin{aligned} \mathbb{1}_E(x) &= 1 \\ &= 0 + 0 + \cdot + 0 + 1 + 0 + \cdots \\ &= \mathbb{1}_{E_1}(x) + \mathbb{1}_{E_2}(x) + \cdots + \mathbb{1}_{E_{n'-1}}(x) + \mathbb{1}_{E_{n'}}(x) + \mathbb{1}_{E_{n'+1}}(x) + \cdots \\ &= \sum_{n=1}^{\infty} \mathbb{1}_{E_n}(x). \end{aligned}$$

Thus, we have established our claim. Thus ,

$$\mathbb{P}(\cup_n E_n) = \mathbb{P}(E) = \int_0^1 \mathbb{1}_E(x)\, dx = \int_0^1 \sum_{n=1}^{\infty} \mathbb{1}_{E_n}(x)\, dx = \sum_{n=1}^{\infty} \int_0^1 \mathbb{1}_{E_n}(x)\, dx = \sum_{n=1}^{\infty} \mathbb{P}(E_n)$$

which is precisely countable additivity[b]. □

All together, we conclude that $\mathbb{P}$ defined in this way is a probability measure. For us, it will model the continuous game of Plinko in such a way that the probabilities of the marble landing in intervals of equal lengths are equally likely. As we will see later, for this reason, we shall append the word "uniform" to such measures on uncountable sample spaces.

[a] In fact, this measure is called the Lebesgue measure and named after French mathematician Henri Lebesgue. The measure's existence is a completely non-trivial business and it isn't defined on all subsets of $[0, 1]$, only those that are so-called "Lebesgue measurable" – which we will take synonymous with "Events". We won't need to worry about any of this in this course, but you should know that there is some really sophisticated mathematical machinery going on here. The study of this is called "measure theory" and is often the subject of MA439.

[b] Note: In the above "proof", an essential piece of my argument relied on exchanging the summation and the integral (i.e., $\int \sum = \sum \int$). Though this property may seem obvious – and it works in this case by something called the monotone convergence theorem – it does not hold in general. So, one needs more robust machinery and it's actually better just to argue directly.

Note Here

Exercise 2.4: Throw a marble in a box

Let's suppose that we have a large box of dimensions 4×6 (linear measurements given in feet). We want to play a game that consists of throwing a marble in the box and seeing where it lands. We can represent the sample space by the rectangle

$$\Omega = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 4 \,\text{and}\, 0 \leq y \leq 6\}$$

where (x, y) represents the landing position of the marble.

1. Describe (explicitely) the following events:

   E_1 : The marble lands within 1 foot of the center of the box.

   E_2 : The marble lands within 1 foot of the bottom right corner of the box.

   E_3 : The marble lands within the top half of the box.

2. Let's now think about assigning a probability measure $\mathbb{P}$ to Ω. There are many ways to assign this $\mathbb{P}$ but, by contrast to what you might expect, an "obvious" one is problematic – this is the assignment of all points to be equally likely. To see this, suppose that all singleton events have the same probability. That is, assume that there is a positive number $p > 0$ such that, for each $(x, y) \in \Omega$,

   $$\mathbb{P}(\{(x, y)\}) = p > 0.$$

   Explain why this $\mathbb{P}$ would violate at least one of the axioms of probability.

3. Due to the shortcoming seen in the previous item, there still is a probability measure $\mathbb{P}$ that makes certain things (in particular, regions of equal area) equally likely. It is defined by

   $$\mathbb{P}(E) = \alpha \times \text{Area}(E)$$

   whenever E is a reasonable[a] event/subset of Ω; here α is a constant. Please use the axioms of probability to determine the constant α. Also, explain why $\mathbb{P}(\{(x, y)\}) = 0$ for each $(x, y) \in \Omega$.

4. With this probability $\mathbb{P}$, compute the probability of the events E_1, E_2, and E_3 you described above.

5. We saw above that each singleton event has 0 probability. Using the axioms of probability, show that every event containing finitely many outcomes also has 0 probability. In other words, if

   $$E = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

   show that $\mathbb{P}(E) = 0$.

6. Can you find an event E containing an infinite collection of outcomes for which $\mathbb{P}(E) = 0$?

[a] Such events are technically called *measurable*.

2.2 Deduction from the Axioms

In this short section, we shall establish some basic properties which follow from the definition of probability measure. We start with a basic lemma which may appear obvious.

Lemma 2.8. *Let $\mathbb{P}$ be a probability measure on Ω. Then $\mathbb{P}(\varnothing) = 0$.*

Proof. Let $E_1, E_2, \ldots,$ be a countable collection of disjoint events; we note that such a collection always exists because the Es can all be taken to simply be the nullevent. Given such a collection, let's form a new collection

of events by setting $F_1 = \varnothing$, $F_2 = E_1$ and, more generally, $F_n = E_{n-1}$ for all $n \geq 2$. This resulting collection of events $F_1, F_2, \ldots$ must be pairwise disjoint because we have only appended the nullevent which doesn't overlap with anything. Also, we have

$$\bigcup_{n=1}^{\infty} E_n = \bigcup_{n=1}^{\infty} F_n.$$

Thus, by countable additivity, we have

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}(E_n) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \\
&= \mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n\right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}(F_n) \\
&= \mathbb{P}(F_1) + \sum_{n=2}^{\infty} \mathbb{P}(F_n) \\
&= \mathbb{P}(\varnothing) + \sum_{n=1}^{\infty} \mathbb{P}(E_n);
\end{aligned}$$

here, we have used the fact that $\mathbb{P}(F_n) = \mathbb{P}(F_{n-1})$ for all $n \geq 2$. In other words, $\mathbb{P}(\varnothing)$ is a number having the property that

$$x = \mathbb{P}(\varnothing) + x$$

for the real number $x = \sum_n \mathbb{P}(E_n)$ and so $\mathbb{P}(\varnothing) = 0$ since zero is the unique number for which this is true. □

Armed with this simple lemma, we are now able to show that countable additivity implies finite (and hence pairwise) additivity. Thus, our second (full) definition of probability measure recaptures the first one. Precisely, we have the following.

Proposition 2.9. *Let $\mathbb{P}$ be a probability measure on a sample space Ω. Then, for any finite collection of disjoint events $E_1, E_2, \ldots, E_N$, we have*

$$\mathbb{P}\left(\bigcup_{n=1}^{N} E_n\right) = \sum_{n=1}^{N} \mathbb{P}(E_n).$$

Proof. Given our finite collection of disjoint events, we shall form a countably infinite collection in the following way: For $n = 1, 2, \ldots, N$, define $F_n = E_n$ and, for $n > N$, set $F_n = \varnothing$. As we argued before, the addition of the nullevent adds nothing to our collection and the new collection remains pairwise disjoint. Also, it is clear that

$$\bigcup_{n=1}^{N} E_n = \bigcup_{n=1}^{\infty} F_n$$

as the nullevents F_n for $n > N$ add nothing to this union. Using this, countable additivity and the the property of

series that we can "peel" of a finite number of terms, we have

$$\begin{aligned}\mathbb{P}\left(\bigcup_{n=1}^{N} E_n\right) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n\right)\\ &= \sum_{n=1}^{\infty}\mathbb{P}(F_n)\\ &= \sum_{n=1}^{N}\mathbb{P}(F_n)+\sum_{n=N+1}^{\infty}\mathbb{P}(F_n)\\ &= \sum_{n=1}^{N}\mathbb{P}(E_n)+\sum_{n=N+1}^{\infty}\mathbb{P}(\varnothing)\\ &= \sum_{n=1}^{N}\mathbb{P}(E_n)\end{aligned}$$

since

$$\sum_{n=N+1}^{\infty}\mathbb{P}(\varnothing)=\sum_{n=N+1}^{\infty}0=0$$

in view of the preceding lemma. □

The following proposition records some basic facts which you were already asked to established with our "trial" definition of probability measure. Since the proposition above shows that our updated (and final) definition of probability measure is countably additive, the results of Exercise 2.3 are valid and we shall simply record these conclusions below without further proof.

Proposition 2.10. *Let $\mathbb{P}$ be a probability measure on a sample space Ω. Then the following statements hold:*

1. *For any events E and F with $E \subseteq F$, we have*

$$\mathbb{P}(E) \leq \mathbb{P}(F).$$

*This is called **monotonicity of measure** or **monotonicity of probability**.*

2. *For any events E and F, we have*

$$\mathbb{P}(E \cup F) \leq \mathbb{P}(E) + \mathbb{P}(F);$$

*this is called **Boole's Inequality**.*

The following proposition extends Boole's inequality to a countable number of events. In modern probability research this inequality is more commonly referred to as **the union bound**.

Theorem 2.11 (The Union Bound)**.** *Let $\mathbb{P}$ be a probability measure on a sample space Ω and let $E_1, E_2, \ldots$ be a countable list of events. Then*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty}\mathbb{P}(E_n).$$

Proof. Our goal is to cook up a collection of disjoint events $F_1, F_2, \ldots$ with the property that $\cup_n F_n = \cup_n E_n$ so that, for the sets $F_1, F_2, \ldots$, we can use the countable additivity of $\mathbb{P}$. To this end, set $F_1 = E_1$, $F_2 = E_2 \setminus E_1$, $F_3 = E_3 \setminus (E_1 \cup E_2)$ and, in general,

$$F_n = E_n \setminus \left(\bigcup_{k=1}^{n-1} E_k\right)$$

for $n \geq 2$. In other words, $F_1 = E_1$, and all the F_ns are defined in a way so that they include what's in E_n but not in any $E_1, E_2, \ldots, E_{n-1}$. To really get a good grasp of how the sets $F_1, F_2, \ldots$ are defined, you should draw some examples. I claim that

$$\bigcup_{n=1}^{\infty} F_n = \bigcup_{n=1}^{\infty} E_n.$$

To see this, first note that $F_n \subseteq E_n$ for each n and so it must hold that

$$\bigcup_{n=1}^{\infty} F_n \subseteq \bigcup_{n=1}^{\infty} E_n. \tag{2.1}$$

It therefore remains to show the reverse containment. To this end, let ω be an (arbitrary) outcome of $\cup_n E_n$. This means that $\omega \in E_n$ for some n. Now, ω could be in many different E_n's, but there must be some "first" event in the list to which it belongs. Let's denote this first event by E_{n_0} where (precisely)

$$n_0 = \min\{n \in \mathbb{N}_+ : \omega \in E_n\}.$$

We have two possible cases: $n_0 = 1$ and $n_0 > 1$. If $n_0 = 1$, then $\omega \in E_1 = F_1 = F_{n_0}$. If instead $n_0 > 1$, by the definition of n_0, ω is in E_{n_0} but it cannot be in E_n for any $1 \leq n \leq n_0 - 1$. Thus,

$$\omega \in E_{n_0} \setminus \left(\bigcup_{n=1}^{n_0-1} E_n \right) = E_{n_0} \setminus \left(\bigcup_{k=1}^{n_0-1} E_k \right) = F_{n_0}.$$

Hence, it always holds that $\omega \in F_{n_0}$ and so

$$\omega \in \bigcup_{n=1}^{\infty} F_{n_0}.$$

We have therefore shown the asserted reverse containment and so (2.1) holds. Furthermore, I claim that the sets $F_1, F_2, \ldots$ are disjoint. To see this, suppose that

$$\omega \in F_n \cap F_k$$

for some $n \geq k$. Let's assume that (without loss of generality) that $n > k$. Then $\omega \in E_k$ since F_k is a subset of E_k. However, by the construction of F_n which doesn't contain the outcomes of E_l for any $l < n$ including E_k, ω cannot belong to E_k. We have arrived at a contradiction – since ω cannot belong and not belong to E_k. Thus, no such outcomes ω exists and hence $F_n \cap F_k = \varnothing$. I have therefore shown that the list $F_1, F_2, \ldots$ is pairwise disjoint.

In view of these properties of the list $F_1, F_2, \ldots$, I am able to apply the countable additivity of $\mathbb{P}$. We have

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n \right) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(F_n)$$

Since $F_n \subseteq E_n$ for all n (by definition), the monotonicity of $\mathbb{P}$ guarantees that

$$\mathbb{P}(F_n) \leq \mathbb{P}(E_n)$$

for all n and hence

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} \mathbb{P}(F_n) \leq \sum_{n=1}^{\infty} \mathbb{P}(E_n)$$

which is our desired inequality. □

Exercise 2.5: Partitions and Probability Measures

Let $\mathbb{P}$ be a probability measure on a sample space Ω.

1. Let $S_1, S_2, \ldots,$ be a partition of Ω. Prove that, for any event E,
$$\mathbb{P}(E) = \sum_{n=1}^{\infty} \mathbb{P}(ES_n).$$
2. Use the previous item to show that
$$\mathbb{P}(E) = \mathbb{P}(EF) + \mathbb{P}(E \setminus F)$$
for any events E and F.

We finish this short section with an important property of probability measures that will allow us to approximate the probabilities of certain events using the probabilities of simpler events. This is known as the **continuity of probability**.

Theorem 2.12 (The Continuity of Probability)**.** *Let $\mathbb{P}$ be a probability measure on a sample space Ω and let $E_1, E_2, \ldots$ be a countably infinite collection of events.*

1. *If the sets $E_1, E_2, \ldots$ are nested increasing in the sense that*
$$E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots,$$
then
$$\mathbb{P}(\overline{E}) = \lim_{n\to\infty} \mathbb{P}(E_n)$$
where
$$\overline{E} = \bigcup_{n=1}^{\infty} E_n.$$
2. *If the events $E_1, E_2, \ldots$ are nested decreasing in the sense that*
$$E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots,$$
then
$$\mathbb{P}(\underline{E}) = \lim_{n\to\infty} \mathbb{P}(E_n)$$
where
$$\underline{E} = \bigcap_{n=1}^{\infty} E_n.$$

Proof. Let's assume that our collection $E_1, E_2, \ldots$ is nested increasing. Set $F_1 = E_1$, $F_2 = E_2 \setminus E_1$ and, in general, $F_n = E_n \setminus E_{n-1}$ for $n > 1$. It is not difficult to show that the events $F_1, F_2, \ldots$ are disjoint and, because the collection of events $E_1, E_2, \ldots,$ is nested increasing, we have
$$\overline{E} = \bigcup_{n=1}^{\infty} E_n = \bigcup_{n=1}^{\infty} F_n.$$
If this isn't clear, I strongly encourage you to draw some pictures and and try to verify these two properties are true. By the countable additivity of $\mathbb{P}$, we have Therefore,
$$\mathbb{P}(\overline{E}) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} F_n = \lim_{N\to\infty} \sum_{n=1}^{N} \mathbb{P}(F_n) = \lim_{N\to\infty} \left(\mathbb{P}(F_1) + \mathbb{P}(F_2) + \cdots + \mathbb{P}(F_N)\right)$$

where I have used the definition that the sum of the series is simply the limit of partial sums. Observe that since $F_1 = E_1$, $\mathbb{P}(F_1) = \mathbb{P}(E_1)$ and, since $F_n = E_n \setminus E_{n-1}$ for $n > 1$, we have $\mathbb{P}(F_n) = \mathbb{P}(E_n) \setminus \mathbb{P}(E_{n-1})$ for $n > 1$. Thus

$$\begin{aligned}
\mathbb{P}(\overline{E}) &= \lim_{N\to\infty} (\mathbb{P}(F_1) + \mathbb{P}(F_2) + \cdots + \mathbb{P}(F_N)) \\
&= \lim_{N\to\infty} (\mathbb{P}(E_1) + (\mathbb{P}(E_2) - \mathbb{P}(E_1)) + \cdots + (\mathbb{P}(E_{N-1} - \mathbb{P}(E_{N-2})) + (\mathbb{P}(E_N) - \mathbb{P}(E_{N-1})) \\
&= \lim_{N\to\infty} (\mathbb{P}(E_1) - \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_2) + \cdots + \mathbb{P}(E_{N-1}) - \mathbb{P}(E_{N-1}) + \mathbb{P}(E_N)) \\
&= \lim_{N\to\infty} (0 + 0 + \cdots + \mathbb{P}(E_N)) \\
&= \lim_{N\to\infty} \mathbb{P}(E_N).
\end{aligned}$$

In other words, our so-called telescoping sum collapses and leaves us with precisely the final term $\mathbb{P}(E_N)$. Noting only that N is a dummy variable, we have shown that

$$\mathbb{P}(\overline{E}) = \lim_{n\to\infty} \mathbb{P}(E_n).$$

This proves the first item, i.e., the continuity of probability for a collection of nested increasing events. The exercise below will show you that the second item can be seen as a quick consequence of the first. □

Exercise 2.6: Nested Increasing to Decreasing

Use De Morgan's Laws and the nested increasing version of the continuity of probability to prove the nested decreasing version of the continuity of probability. Hint: If $E_1, E_2, \ldots$ is nested decreasing, then $E_1^c, E_2^c, \ldots$ is nested increasing.

Exercise 2.7:

Suppose that your professor is late to class and, at time $n = 0$, your professor emails and said that the probability that you have to wait at least n minutes is

$$p_n = \frac{5 + e^n}{3(1 + e^n)}.$$

What is the probability that your professor will never show up?

Chapter 3

How to Count

It is often the case that we are confronted with an experiment whose sample space Ω is finite and the most reasonable probability measure is the uniform measure $\mathbb{P}_u$. For example, consider an experiment where we select five cards from a deck of 52 (without replacement) and ask about the probability of each five-card outcome (called a *hand*). If the selection of cards is done with a uniformly well-shuffled deck and each card is selected independently of those before it, it is reasonable to expect that each hand is equally likely. This means that we should model our probabilities by the uniform measure $\mathbb{P}_u(E) = \#(E)/\#(\Omega)$ on the sample space of five-card hands. Thus, to compute the probabilities of various events (e.g., the probability of having three of a kind or a flush), we must be able to count the number of hands in a given event E as well as the number of total hands, $\#(\Omega)$. In general, this is far from a simple task. As you're reading, you should pause here and see can compute $\#(\Omega)$ and $\#(E)$ where E is the event that your hand has three of a kind.

This chapter is therefore dedicated to establishing some various techniques for counting. Let's begin with some basic questions:

1. If a restaurant's menu has four hamburgers and seven sides (e.g., french fries), how many plates (including a hamburger and side) can I order? *If they also sever 18 types of sodas, how many hamburger-side-soda combos can I order?*
2. If I have 10 books, how many ways can I place them on a shelf? *What if two books are the same?*
3. Of all sequences $(x_1, x_2, \ldots, x_n)$ with $x_j \in \{0, 1\}$ for $j = 1, 2, \ldots, n$, how many sequences contain exactly k ones?

We begin with a fundamental rule of counting:

Principle 3.1 (Fundamental Principle of Counting)**.** *There are N choices to be made. There are m_1 possible choices for the first, m_2 choices for the second, m_3 for the third,... and m_N for the Nth. If these choices are to be combined freely (no choice affects the others), the total number of choices for the set is*

$$m_1 m_2 \cdots m_N.$$

Note Here

With this principle, we can immediately answer the first question above.

Example 3.1: Hamburger and a Side

If a restaurant menu has four hamburgers and seven sides, we can think of a plate as a choice of one burger and one side. Given that there are four burgers and seven sides, we have $4 \times 7 = 28$ possible plates to be ordered. If we include the possibility of also choosing one of 18 sodas, we have $4 \times 7 \times 18 = 28 \times 18 = 504$ possible hamburger-side-soda combos that can be made.

In looking to answer the second question above, i.e., that in which we place books on a shelf, we notice that after a first book is chosen, there are fewer remaining books to choose from. In other words, the first choice does affect the second (and third and fourth). Hence, we cannot directly apply the fundamental principle of counting. However, as we shall see, the principle does provide a key to answer the question. For now, let's focus on some strategies.

3.1 Different Ways to Sample

Many of the counting problems that we shall reduce to questions of sampling. We can think of sampling as a process by which we have a number of things to choose from – we shall call them balls in an urn[1]. Let's suppose then that there are n balls in an urn and, at least at first, each ball gets a number from 1 to n. We will then begin to select balls from the urn; this is called sampling. We focus on four main scenario:

1. With replacement and with order.
2. Without replacement and with order.
3. Without replacement and without order.
4. With replacement and without order.

3.1.1 Sampling With Replacement and With Order

In this scenario, let's choose a ball from the urn, mark down the number that appears on its face (from 1 to n) and then replace the ball in the urn. We then repeat this m times. We can ask: How many possible ways are there for this to happen? Thinking very carefully, we see this question is exactly the same as:

How many sequences $(x_1, x_2, \ldots, x_m)$ can be made where each x_j can be chosen from the set $\{1, 2, \ldots, n\}$?

For these questions, we see that each choice has the full range of possibilities and is therefore done freely of the other choices. In other words, for each selection of a ball from the urn, the entire set of balls is available to choose from; this is precisely because the balls are replaced each time and so we say that this sampling is done **with replacement**. In terms the sequence, this means that the entire range of $\{1, 2, 3, \ldots, n\}$ is available for each entry x_j. Also, because we are marking down the choice of balls in an order (and the entries of the sequence also have an order), we say that this sampling is done **with order**.

Answering this question boils down exactly to the fundamental principle of counting because our choices are made freely and independently of each other. Since there are m choices made and each choice has n options, we have

$$n^m = \underbrace{n \cdot n \cdot n \cdots n}_{m}$$

possible selection of balls and, equivalently, n^m possible sequences.

3.1.2 Sampling Without Replacement and With Order

Like the last scenario, we choose balls consecutively from the run, however, after each choice we do not place the ball back into the urn. In other words, we choose a ball from the urn and set it aside. We do this m times (with necessarily $m \leq n$). Since we are not putting the balls back into the urn, we call this sampling **without replacement**. Still, we are paying attention to the order in which the balls were drawn and so we are still in the scenario in which this sampling is done with order.

[1] In most treatments of probability (at every level), you will see a focus of choosing things from urns. You might ask: why not bags or bushels, or bowls? The answer seems to date back to one of the progenitors of the modern theory of probability, Jacob Bernoulli. Bernoulli's famous treatise on the subject, *Ars Conjectandi* from 1713, introduces the notion of selecting marbles from an urn (and discussing the probability of such experiments). Today, these questions, there is an entire area of probability and statistical mechanics known as "urn models".

I like to think of this scenario as one in which we have a list of books from which we can choose and place on a shelf. Here, the order in which the books appear does matter and so our question of how many ways in which there is to choose m balls from an urn of n balls is the same as asking the number of ways that we can arrange m books on a shelf where the books are chosen from a lot of n.

To solve this problem, it is helpful to think about each book (or slot) at a time. For the first, choice there are n books to choose from and so we count

$$\underbrace{n\_\_\_\_\cdots\_\_}_{m}$$

as the remaining slots have not yet been filled. For our second choice, we have $n-1$ books to choose from and so we count

$$\underbrace{n\ n-1\_\_\_\cdots\_\_}_{m}.$$

We continue doing this until we have made our m choices and in this way we count

$$\underbrace{n\,n-1\,n-2\cdots n-(m-1)}_{m}.$$

Thinking about each filled slot as the number of choices that could be made in that slot, the number of total ways (e.g., selection of balls without replacement and with order or books on a shelf) is simply the product of the numbers appearing in these slots. This is the number

$$n(n-1)(n-2)\cdots(n-(m-1)).$$

If we think in terms of factorials, observe that

$$\begin{aligned}\frac{n!}{(n-m)!} &= \frac{n(n-1)(n-2)\cdots(n-(m-1))(n-m)(n-m-1)\cdots(2)(1)}{(n-m)(n-m-1)\cdots(2)(1)}\\ &= n(n-1)(n-2)\cdots(n-(m-1))\end{aligned}$$

which is precisely the same number. Hence there are

$$\frac{n!}{(n-m)!}$$

possible ways to choose m balls from an urn of n balls without replacement and with order.

Example 3.2: The lottery

A wheel of ninety nine balls numbered $\{1,2,\ldots,98,99\}$ spins mixing up the balls. A lottery number is drawn by selecting five balls from the wheel consecutively without replacement. How many lottery numbers are possible?

Since this is precisely the scenario of sampling $m=5$ balls from $n=99$ without replacement but with order (lottery numbers are ordered), we have

$$\frac{99!}{(99-5)!}=\frac{99!}{94!}=(99)(98)(97)(96)(95)=8582777280$$

possible lottery numbers.

There is a special case of sampling without replacement and with order that is important enough to highlight. Let's suppose that we are tasked with counting the number of sequences of $m=n$ balls that can be drawn from the urn (or n) balls. We can see this equivalently as the number of **permutations** or **orderings** of n things. With the prescription above, this number is simply $n!/(n-n)!=n!/0!=n!$ where we recall[2] that $0!=1$. Thus, the number of permutations of n things is exactly $n!$.

[2]Have you thought about why $0!=1$? After all, it is a convention but, conveniently, it does make valid the identity $n!=n(n-1)!$ for all $n=1,2,\ldots$.

3.1.3 Sampling Without Replacement and Without Order

In this scenario, we draw m balls from the list of $\{1, 2, \ldots, n\}$ and, after each choice, we do not replace the ball. In contrast to the last scenario, we do not pay attention to the order in which the balls were drawn. To treat this scenario, let's consider an example.

Example 3.3: Getting Rid of Order

Let's choose $m = 3$ balls from a list of $n = 5$ without replacement. At first, let's pay attention to the order of the balls so that we fall into the previous scenario in which there are $5!/(5-3)! = 5!/2! = 60$ possible different sequences. Consider, for example the following sequences in which the balls 1, 3, and 4 were all chosen:

1, 3, 4	1, 4, 3
3, 1, 4	3, 4, 1
4, 1, 3	4, 3, 1

We note that there are precisely 6 of these sequences because there are $3! = 6$ ways to arrange 3 distinct things in 3 slots; this is the number of permutations of 3 balls.

Let's now stop paying attention to the order in which the balls are drawn. In this situation, we would regard the $3! = 6$ sequences above as the same and, in fact, for each choice of 3 distinct balls, it is clear that are are $3! = 6$ distinct sequences that can be made from the three balls. Thus, the $5!/2! = 60$ distinct sequences can be divided into 10 groups of $3! = 6$ where each groups contains all sequences containing the same three balls. Since we're not paying attention to order, we simply regard each group as the same and we count that there are

$$10 = \frac{5!}{3!2!} = \frac{5!}{(5-2)!2!} = \frac{n!}{(n-m)!m!}$$

groups of $m = 3$ balls that can be made from $n = 5$.

The example above illustrates exactly how we can count the number of distinct groups of m balls drawn from the urn containing n balls. We start first by paying attention to order (so we are in the previous scenario) and count that there are

$$\frac{n!}{(n-m)!}$$

distinct sequences that can be made with m balls from the urn. For any m distinct balls, there are exactly $m!$ sequences in this list of distinct sequences which contain exactly these m balls – this is the number or permutations of m balls. Thus, we can divide the original list of sequences up into groups containing the same m balls and, in the present scenario in which we do not pay attention to order, the number of distinct collections of m elements is exactly the number of groups formed. This is

$$\frac{n!}{(n-m)!m!}.$$

As we will see, this number is important enough to give a name and notation to. For non-negative integers n and m with $m \leq n$, the **binomial coefficient** n **choose** m is the number

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}$$

which is precisely the number of groups/collections of m things that can be formed with n things (without replacement). If you have not seen it before, the reason for calling these "binomial coefficients" will soon be made clear.

This is another case of sampling without replacement and without order that's worth mentioning. Let's suppose that we choose n balls from an urn of n balls. Instead of the n balls having distinct labeling, let's suppose that there are r subcollections of these n balls where the balls in each subcollection are indistinguishable. In other

words, given a set of n items, suppose that there are r subsets containing (respectively) $n_1, n_2, \ldots, n_r$ items (with $n = n_1 + n_2 + \cdots + n_r$) where the elements/items in each of these subsets are indistinguishable. In this case, the number of ways that we can order all n items given that we cannot distinguish between permutations of the items within the subsets is

$$\frac{n!}{n_1!n_2!\cdots n_r!}.$$

The following example illustrates this.

Example 3.4: Apples, Oranges, and Bananas

Let's suppose that we have a bushel containing $m_1 = 3$ apples, $m_2 = 2$ oranges, and $m_3 = 2$ bananas and we'd like to distribute the content of the bushel to $m = 3+2+2 = 7$ hungry students. How many ways can this happen?

To solve this, let's first suppose that we can tell the difference between each apple in the collection of 3, orange in the collection of 2, and banana in the collection of 2. Here, we might as well give them the labeling

$$\{A_1, A_2, A_3, O_1, O_2, B_1, B_2\}.$$

Since they are (at present) all distinguishable, there are clearly $n! = 7!$ ways to distribute the fruit to the 7 students; this is the number of permutations of 7 things. Let's consider six different such ways (here, I'll keep track of order simply by juxtaposition):

$$\begin{aligned} &A_1B_1O_1A_2A_3O_2B_2 \qquad A_1B_1O_1A_3A_2O_2B_2\\ &A_2B_1O_1A_1A_3O_2B_2 \qquad A_2B_1O_1A_3A_1O_2B_2\\ &A_3B_1O_1A_1A_2O_2B_2 \qquad A_3B_1O_1A_2A_1O_2B_2 \end{aligned}$$

If one could not distinguish between A_1, A_2 and A_3, we would regard each of these as the same. In fact, for each permutation of O_1 and O_2 and B_1, and B_2 where the O's appear in the third and sixth slots and the Bs appear in the second and seventh slots, we see that there are $3! = 6$ different arrangements of A_1, A_2, and A_3 that should be recognized as the same. By grouping such sequences together, we see that $n! = 7!$ over counts (multiplicatively) by 3!, the number of permutations of 3 apples. Thus, we are left with

$$\frac{n!}{n_1!} = \frac{7!}{3!} = 120$$

sequences made by

$$\{A, A, A, O_1, O_2, B_1, B_2\}.$$

Repeating this line of argument again, we that if we do not distinguish between oranges, we can group according to permutations of the O's. This again shows that we have overcounted multiplicatively by $n_2! = 2!$ and so we find that the number of distinct sequences that can be made by

$$\{A, A, A, O, O, B_1, B_2\}$$

is

$$\frac{n!}{n_1!n_2!} = \frac{7!}{3!2!} = 60.$$

We then repeat the argument one final time to made indistinguisable B_1 and B_2 and we find that there are

$$\frac{n!}{n_1!n_2!n_3!} = \frac{7!}{3!2!2!} = 30$$

distinct sequences that can be formed with

$$\{A, A, A, O, O, B, B\}$$

which is precisely the number of ways that the bushel of fruit can be distributed among the students.

3.1.4 Sampling With Replacement and Without Order: *Stars and Bars*

Let's consider one final scenario. Given our urn containing n balls, let's select m balls with replacement but without order. To illustrate a general method for counting in this scenario, let's consider an example with $m = 3$ and $n = 4$.

Example 3.5: Balls from an urn

Let's select $m = 3$ balls from an urn containing $n = 4$ balls with replacement but without order. Since we are sampling without order, we can keep track of each distinct set by writing down the unique corresponding non-increasing sequence for our choice of 3 balls. For example, if our selection yields the balls $\{1, 3, 1\}$, we can indicate this by writing the non-increasing sequence $1, 1, 3$. With this presentation, let's consider some of the possibilities with the hopes of establishing a pattern:

Seq.	1	2	3	4
1,1,1	$\star\star\star$			
1,1,2	$\star\star$	$\star$		
1,2,3	$\star$	$\star$	$\star$	
3,3,4			$\star\star$	$\star$

In the table above, I have listed several possible sequences and, for every appearance of a number $k = 1, 2, 3, 4$, I've marked a star in the kth column. For example, the sequence $1, 1, 2$ has two stars in the first column, one star in the second column and no stars in the third or fourth columns. In looking at it this way, it isn't terribly surprising that the placement of stars among columns (including leaving columns empty of stars) captures each sequences completely. In the following table, I have added a final column labeled "pattern" that aggregates this information.

Seq.	1	2	3	4	Pattern
1,1,1	$\star\star\star$				$\star\star\star\|\|\|$
1,1,2	$\star\star$	$\star$			$\star\star\|\star\|\|$
1,2,3	$\star$	$\star$	$\star$		$\star\|\star\|\star\|$
3,3,4			$\star\star$	$\star$	$\|\|\star\star\|\star$

The starts and bars pattern in the table above gives us the key to count in the present scenario (with replacement and without order). Our task is to then count how many ways we can arrange $m = 3$ stars among $n - 1 = 4 - 1 = 3$ bars (note, we only need to pay attention to the bars separating numbers which is why we have the -1). Thinking about it this way, we can approach the counting as we did in the previous scenario wherein we counted permutations of objects which could be broken up into r subcollections where the objects in each subcollection are indistinguishable. In total, we have $m+n-1 = 6$ objects which appear in two subcollections (stars m and bars $n - 1$) and so we have

$$\frac{(m+n-1)!}{m!(n-1)!} = \frac{6!}{3!3!} = 40$$

total possible patters/sequences/ways.

The example above gives outlines a general prescription for counting the total number of possible selections of choosing m balls from an urn of n balls with replacement but without order. In thinking of this general situation as counting the number of patterns, the question becomes: How many ways can we arrange m stars among $n - 1$ bars. This number is

$$\frac{(m+n-1)!}{m!(n-1)!}$$

and it is given (equivalently) by the binomial coefficients

$$\binom{m+n-1}{m} = \binom{m+n-1}{n-1} = \frac{(m+n-1)!}{m!(n-1)!}.$$

3.2 Some Examples

In this section, we work through several examples which illustrate and reinforce our methods of counting.

Example 3.6: Four Queens in a Row

Let's assume that we have a well-shuffled standard deck of 52 cards and we consecutively turn over all cards. We ask: What is the probability that we find four queens in a row?

In this situation, assumption of the well-shuffled deck indicates that our model for probability ought to be done with the uniform measure. Let's first focus on describing the sample space Ω. Since we are selecting cards in order, this is a situation in which we are sampling without replacement but with order. We see that Ω can be described as the collection of sequences of length 52 selected from $\{1, 2, \ldots, 52\}$ without replacement. Thus, we have

$$\#(\Omega) = \underline{1}\ \underline{2}\ \underline{3} \ldots \underline{51}\ \underline{52} = 52!.$$

Now, let's focus on the event E in which we find all Queens in a row in our selection of all 52 cards. We can break up our event E by paying attention to the number of selections before the first Queen is pulled. We have

$$E = \bigcup_{k=1}^{49} E_k$$

where E_k is the event that we obtain 4 Queens in a row starting at the kth selection. For example, E_1 is the event that we first select 4 Queens in a row and then the other non-Queen cards. Similarly, E_{49} is the event that the first 48 selections are non-Queen cards and then the final 4 are Queens. If we first count the order in which the different Queens appear (say we label them Q_1, Q_2, Q_2, Q_4), we see that the number of ways we see the sequence

$$\underline{Q_1}\ \underline{Q_2}\ \underline{Q_3}\ \underline{Q_4}\ \underline{?}\ \underline{?} \cdots \underline{?}$$

is 48!. If we account for the number of ways that the Queens can be arranged (which is 4!), we find that

$$\#(E_1) = 4!48!.$$

In applying this argument to the other various E_ks, we see that

$$\#(E_k) = 4!48!$$

for all $k = 1, 2, \ldots, 49$ and, since the events are disjoint, we have

$$\#(E) = \sum_{k=1}^{49} \#(E_k) = 49(4!48!) = 4!(4948!) = 4!49!.$$

With the uniform measure $\mathbb{P}$, we see that

$$\mathbb{P}(E) = \frac{\#(E)}{\#(\Omega)} = \frac{4!49!}{52!} = \frac{24}{52 * 51 * 50} \approx 0.00018.$$

Example 3.7: Probability of a Pair (or better!)

Let's consider a standard and well-shuffled deck of cards and we will choose 5 cards from the deck to form a standard poker hand. We ask: What is the probability that we have a pair or better[a]?

Because we are working with a well-shuffled deck of cards, we assume that all hands of 5 cards are equally likely and so we shall model this situation using the uniform probability measure. In other words, our task is to count. Frequently, you will find that it is easier to count the outcomes in the complement of an event than it is to count the outcomes in the event itself. This will often be the case when the event is phrased in terms of "this or better" or "at least". In this case, our event is the event E that we obtain a pair or better and so its complement is the event E^c that, in the five cards drawn, there are no pairs. So, our goal is to compute $\mathbb{P}(E^c)$ for then we have $\mathbb{P}(E) = 1 - \mathbb{P}(E^c)$ where $\mathbb{P}$ is the uniform probability measure. Before counting E, let's discuss the sample space Ω. Thinking in terms of drawing the cards one after another (i.e., paying attention to the order in which the cards are drawn), we can represent the sample space as the collection of sequences of five cards $C_1, C_2, \ldots, C_5$ drawn from 52 without replacement but with order. Mathematically, we could write

$$\Omega = \{(C_1, C_2, C_3, C_4, C_5) : C_k \in \{1, 2, \ldots, 52\} \text{ for } k = 1, 2, 3, 4, 5 \text{ and } C_k \neq C_k \text{ when } j \neq k\}.$$

Of course, counting Ω is exactly what we studied in Subsection 3.1.2 and we find that

$$\#(\Omega) = \frac{52!}{(52-5)!} = 52 * 51 * 50 * 49 * 48.$$

Now, to count the event E^c, we can think also in terms of counting without replacement but with order. In choosing the first card, we have 52 options (because we can't have a pair of cards if we only have one card). Once that first card is selected, to make sure we don't make a pair upon the selection of our second card, we have only 48 possible choices for the second card; these are all of the cards that do not share the same face value as the first card selected. By a similar argument, in choosing the third card, we have only 44 possible options if we are not to make a pair using one of the first two cards selected. Continuing in this way, we see that

$$\#(E^c) = 52 * 48 * 44 * 40 * 36$$

and hence

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - \frac{\#(E^c)}{\#(\Omega)} = 1 - \frac{52 * 48 * 44 * 40 * 36}{52 * 51 * 50 * 49 * 48} = \frac{2053}{4165} \approx 0.4929.$$

In the above computation, we paid attention to the order in which the cards were pulled (e.g, in looking at things as sequences). There is another way to count in which we do not pay attention to order. This, of course, necessitates us to represent the sample space a different way. Instead of sequences, our sample space will consists of sets of 5 cards, i.e.,

$$\Omega = \{\{C_1, C_2, C_3, C_4, C_5\} : C_k \in \{1, 2, \ldots, 52\} \text{ for } k = 1, 2, 3, 4, 5 \text{ and } C_k \neq C_k \text{ when } j \neq k\}.$$

To count Ω, we simply choose 5 cards from 52 without replacement and thus

$$\#(\Omega) = \binom{52}{5}.$$

Our event E^c consisting of all sets of 5 cards where every card has a different face value is slightly more difficult to count. Imagine that the deck of cards is divided up into 13 piles where each pile represents a different face value. Then, to count E^c, we can first choose 5 distinct piles from the 13 (this is the situation in which we cannot have a pair or more) and then, for each of the 5 piles, we have a choice of 4 cards, one for each suit. Thinking of this as a two-step procedure, we have

$$\#(E^c) = \binom{13}{5} 4^5$$

where the binomial coefficient represents the choices of piles and the 4^5 represents the number of choices within the piles, 4 from each. Thus

$$\mathbb{P}(E) = 1 - \mathbb{P}(E^c) = 1 - \frac{\binom{13}{5}4^5}{\binom{52}{5}} = \frac{2053}{4165},$$

as it must be. This example illustrates that there are often many different ways to arrive at a correct answer. What's important is that, for any approach, you count both the event and the sample space consistently. For example, if you pay attention to order in your counting of an event, you must also pay attention to this in the counting of the sample space.

[a] Better means that we could have three or four of a kind, or two pair, or two pair and three of a kind. For this game, we are ignoring other desirable hands like flushes.

Example 3.8: The Birthday Problem

In a room of n people, what is the probability that at least two people share a birthday?

To give a reasonable answer to this question, we must first make some assumptions to form a probability model. The first assumption we shall make is that no person in the room was born on February 29th of a leap year. With this, we shall also assume that each person is equally likely to be born on any of the 365 days of the year. Though being born on February 29th of a leap year has a non-zero probability, the probability is actually quite small (≈ 0.0007) and so this assumption is quite reasonable. It is less reasonable to assume that all 365 of the remaining days in any year are equally likely. For example, children are more likely to be conceived in winter [6]. In any case, our assumptions above will provide a reasonable model of probability that is testable. We assume that each person is equally likely to be born on any day of the 365-day year and hence we are working with the uniform probability measure.

The event E_n that there are at least two people who share a birthday is quite difficult to count. As in the previous example, let's instead count the outcomes of the event E_n^c that none of the n people share a birthday. In this scenario, we can view our sample space as all sequences of length n where each entry is selected from $1, 2, \ldots, 365$. Thus,

$$\#(\Omega_n) = (365)^n$$

and we note that this representation (and count) were done paying attention to order. We must therefore pay attention to order when we count E_n^c. Counting E_n^c is simple – it is just the scenario of Subsection 3.1.2 where we select n days consecutively from 365 without replacement (because we're asking for each day to be different). Hence

$$\#(E_n^c) = \frac{(365)!}{(365-n)!}$$

and so

$$\mathbb{P}(E_n) = 1 - \frac{(365)!}{(365)^n(365-n)!}.$$

Let's compute this for a couple of small n cases. First, notice that $\mathbb{P}(E_1) = 1 - 1 = 0$ which is expected since we cannot have two (or more) shared birthdays in a room of n people. $\mathbb{P}(E_2) = 1 - (364/365) = 1/365$ which is expected because each day of the year was assumed to be equally likely and so, in a room of two people, the probability of them sharing the same birthday is simply $1/365$. In Figure 3.1, I have plotted this function of n for $1 \leq n \leq 40$.

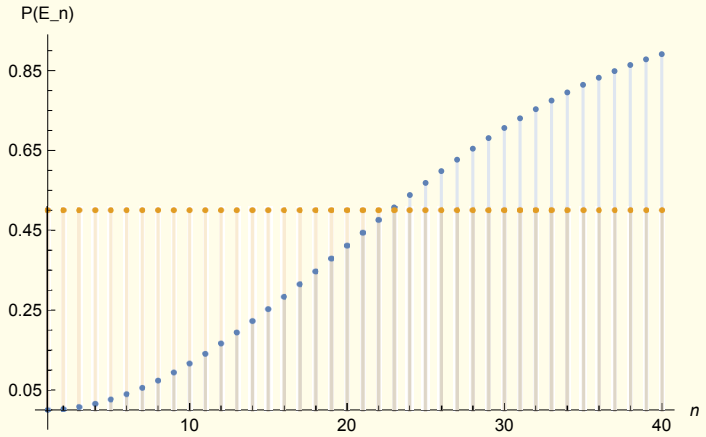


Figure 3.1: Graph of $\mathbb{P}(E_n)$ vs. n.

In looking at the figure, it is remarkable that the probabilities significantly increase around $n = 10$ and, in fact, surpass $1/2$ at $n = 23$. Here,

$$\mathbb{P}(E_{23}) \approx 0.507.$$

This is testable! If you're in a room with n people or more, you can sample people to see if any two share a birthday. You'll find that this model is pretty accurate with it being more probable than not that two people share a birthday in a room of $n \geq 23$. For an interesting generalization of this question and problem, I encourage you to read the article Triple birthday matches in the Senate: Lies, damned lies, and chatGPT" by Rick Durrett [5].

3.2.1 Some Exercises

Exercise 3.1: Basic Counting

Suppose I have $n = 10$ books.

1. How many ways can I arrange 4 of them on a shelf?
2. If, in that 10 books, 3 are identical linear algebra books and 2 are identical probability books, how many ways can I arrange all 10 books?
3. Assuming, as in the previous item, 3 are identical linear algebra books and 2 are identical probability books, how many ways can I arrange 6 of the books given that all of the linear algebra and all of the probability books must appear?

Exercise 3.2: Throwing Dice

In this problem, you will throw m six-sided dice and count the number of possibilities in three different situations.[a]

1. Suppose that the m dice are distinguishable. For instance, suppose that they are all regular six-sided dice but each die is a different color. If you select each die randomly and roll it, how many distinct m rolls can be produced? Hint: You can think of this as a two-step procedure, where you first choose one die out of a bag, roll it and then record both the color and the result of the roll. Not replacing the dice, repeat m times. How many different possibilities are there?

2. Now, suppose that the dice are indistinguishable, and you roll all m and keep track of the order in which the results came up. How many possibilities are there? Hint: In this case, you are counting the number of sequences $(a_1, a_2, \ldots, a_m)$ where the a_k can take on any value from 1 to 6.

3. Suppose that the dice are indistinguishable and you roll m of them at once. How many distinct results are possible?

[a] As with any type of problem where variables are used, if you first find it difficult to think about m dice, try doing the problem for a fixed number, say $m = 5$. Then think about $m = 6$. If you can recognize the pattern, it will help you to solve the problem for a general m.

Exercise 3.3: Choosing Shoes

As with many problems in probability, it is often easier to compute the probability of an event A by first computing $\mathbb{P}(A^c)$ (which might be easier to do) and then concluding that $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$.

1. There are 8 pairs of shoes and you select 4 shoes uniformly at random from the collection of 16 possible shoes. What is the probability that you have selected at least one pair?

2. Generalize this problem: There are n pair of shoes and you select $m \leq 2n$ shoes from the collection. What is the probability that you have at least one pair?

Exercise 3.4: Binomial Coefficients

For any $1 \leq k \leq n-1$, show that

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

Illustrate how this identity is the basic idea behind Pascal's triangle.

Exercise 3.5: How Many Partitions

Consider the sample space

$$\Omega = \{1, 2, \ldots, n\}.$$

In this exercise, we shall focus on counting the number of partitions of Ω where we ask that all events in each partition are non-empty. Denote by T_n the number of such partitions of Ω. Here, we see immediately that $T_1 = 1$ because $\{\{1\}\}$ is the only partition of $S = \{1\}$ and $T_2 = 2$ because there are exactly two partitions of $\Omega = \{1, 2\}$: $\{\{1, 2\}\}$ and $\{\{1\}, \{2\}\}$.

1. By writing down all possible partitions, show that $T_3 = 5$ and $T_4 = 15$.

2. Show that

$$T_{n+1} = 1 + \sum_{j=1}^{n} \binom{n}{j} T_j$$

and use this to compute T_{10}.

Exercise 3.6: Same Birth Month?

In a room of n people, what is the probability that at least two will share the same birth month? You may assume that it is equally likely to be born on any given month. What is the smallest value of n for which this probability is more than $1/2$?

Challenge problem: What is the probability that at least three will be born in the same birthday month?

Exercise 3.7:

Let's suppose that we have an urn containing 5 red marbles and 3 blue marbles. If I select 5 marbles and line them up on my chalkboard rail, what is the probability that the marbles alternate in color?

Exercise 3.8:

Six fair dice are rolled. What is the probability of getting three pairs?

3.3 The Binomial Theorem

For real numbers x and y and a positive natural number n, the binomial theorem gives a formula for the expansion of $(x+y)^n$. Though there are several ways of looking at this theorem, let's try to motivate it through the lens of counting. To get started, for real numbers x and y, we can expand

$$\begin{aligned}(x+y)^2 &= (x+y)(x+y)\\ &= xx+xy+yx+yy\\ &= xx+(xy+yx)+yy\\ &= x^2+2xy+y^2\end{aligned}$$

and

$$\begin{aligned}(x+y)^3 &= (x+y)(x+y)(x+y)\\ &= (x+y)(xx+xy+yx+yy)\\ &= xxx+xxy+xyx+xyy+yxx+yxy+yyx+yyy\\ &= xxx+(xxy+xyx+yxx)+(xyy+yxy+yyx)+yyy\\ &= x^3+3x^2y+3xy^2+y^3.\end{aligned}$$

In the penultimate lines for each expansion above, I have grouped all terms together that contain the same number of xs and ys. Doing this for $n=4$, we have

$$\begin{aligned}(x+y)^4 &= (x+y)(xxx+xxy+xyx+xyy+yxx+yxy+yyx+yyy)\\ &= xxxx+xxxy+xxyx+xxyy+xyxx+xyxy+xyyx\\ &\qquad +xyyy+yxxx+yxxy+yxyx+yxyy+yyxx+yyxy+yyyx+yyyy\\ &= xxxx+(xxxy+xxyx+xyxx+yxxx)+(xxyy+xyxy+xyyx+yxxy+yxyx+yyxx)\\ &\qquad +(xyyy+yxyy+yyxy+yyyx)+yyyy\\ &= x^4+4x^3y+6x^2y^2+4xy^3+y^4.\end{aligned}$$

Continuing this type of pattern (I promise I won't write out $n \geq 5$), the expansion of $(x+y)^n$ can be expressed as a summation of all terms of the form

$$a_1a_2\cdots a_n$$

where a_j is either x or y for $j=1,2,\ldots n$. Since multiplication is commutative for real numbers, we aggregate all such terms in the expansion that have the same number of xs and ys and "collapse" them by writing them (equivalently) as

$$x^{n-k}y^k$$

for $k = 0, 1, \ldots, n$. Of course, to get the correct expansion, we must keep track of how many terms were collapsed and are now being presented as $x^{n-k}x^k$. For example, in the $n = 4$ case, the four terms $xxxy$, $xxyx$, $xyxx$, and $yxxx$ were collapsed into x^3y giving us the final term $4x^3y$ in the expansion above. Thus, for a fixed $k = 0, 1, 2, \ldots, n$, we ask:

How many terms of the form $a_1a_2\cdots a_n$ can be formed with exactly $n-k$ xs and k ys?

This is, of course, a counting problem. Thinking of it in terms of selecting n balls from an urn of n balls where the n balls are divided into two subcollections of indistinguishable balls ($n_1 = n-k$ and $n_2 = k$), we simply count the number of resulting permutations. As we discussed in Subsection 3.1.3, for each k there are

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

permutations. Making sure we keep track of all such permutations and exhaust all possibilities of k, we should therefore have

$$(x+y)^n = \binom{n}{0}x^ny^0 + \binom{n}{1}x^{n-1}y^1 + \cdots + \binom{n}{n-1}x^1y^{n-1} + \binom{n}{n}x^0y^n.$$

In this way, we are led to the binomial theorem.

Theorem 3.2 (The Binomial Theorem)**.** *For any natural number n and real numbers x and y, we have*

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

Though we have motivated the theorem via counting, let's give an inductive proof of the theorem.

Proof. Let x and y be arbitrary but fixed real numbers. Observe that

$$(x+y)^1 = x+y = \binom{1}{0}x^{1-0}y^0 + \binom{1}{1}x^{1-1}y^1 = \sum_{k=0}^{1}\binom{1}{k}x^{1-k}y^k$$

and so the statement holds for $n = 1$. Let's assume that our statement holds for $n-1$ (for some $n \geq 2$, this is the inductive hypothesis) and we shall show it for a general n. We have

$$\begin{aligned}
(x+y)^n &= (x+y)(x+y)^{n-1} \\
&= (x+y)\left(\sum_{k=0}^{n-1}\binom{n-1}{k}x^{n-1-k}y^k\right) \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k}xx^{n-1-k}y^k + \sum_{k=0}^{n-1}yx^{n-1-k}y^k \\
&= \sum_{k=0}^{n-1}\binom{n-1}{k}x^{n-k}y^k + \sum_{k=0}^{n-1}\binom{n-1}{k}x^{n-1-k}y^{k+1}
\end{aligned}$$

For the first summation, we peel off the first term to see that

$$\sum_{k=0}^{n-1}\binom{n-1}{k}x^{n-k}y^k = x^n + \sum_{k=1}^{n-1}\binom{n-1}{k}x^{n-k}y^k = \binom{n}{0}x^{n-0}y^0 + \sum_{k=1}^{n-1}\binom{n-1}{k}x^{n-k}y^k. \tag{3.1}$$

For the second term, we peel of the last term and re-index $k' = k+1$ to see that

$$\begin{aligned}
\sum_{k=0}^{n-1}\binom{n-1}{k}x^{n-1-k}y^{k+1} &= \sum_{k=0}^{n-2}\binom{n-1}{k}x^{n-(k+1)}y^{k+1} + \binom{n}{n}x^{n-n}y^n \\
&= \sum_{k'=1}^{n-1}\binom{n-1}{k'-1}x^{n-k'}y^{k'} + \binom{n}{n}x^{n-n}y^n
\end{aligned}$$

where we note that the summation indices went from $0 \leq k \leq n-2$ to $1 \leq k' \leq n-1$ since $k' = k+1$. Since, k' is a dummy variable of summation, we can write this summation out by dropping the prime and so we have

$$\sum_{k=0}^{n-1} \binom{n-1}{k} x^{n-1-k} y^{k+1} = \sum_{k=1}^{n-1} \binom{n-1}{k-1} x^{n-k} y^k + \binom{n}{n} x^{n-n} y^n. \tag{3.2}$$

Combining (3.1) and (3.2) into our orignal formula for $(x+y)^n$ gives

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

Using the result of Exercise 3.4, we have

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

for $1 \leq k \leq n-1$ and so

$$\begin{aligned}(x+y)^n &= \binom{n}{0} x^{n-0} y^0 + \sum_{k=1}^{n-1} \binom{n}{k} x^{n-k} y^k + \binom{n}{n} x^{n-n} y^n \\ &= \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k\end{aligned}$$

which is precisely the formula for n. □

Throughout the semester, we will see applications of the binomial theorem pop up all over the place, sometimes in unexpected ways. For the moment, let's see a simple and straightforward application.

Example 3.9: Tossing a Fair Coin: How Many Heads?

Let's suppose that we toss a fair coin n times and count the number of times that "heads" appears. As we have seen before, a good representation of the sample space is

$$\Omega_n = \{(x_1, x_2, \ldots, x_n) : x_k \in \{H, T\} \text{ for } k = 1, 2, \ldots n\}$$

where each outcome $\omega = (x_1, x_2, \ldots, x_n)$ has probability $\mathbb{P}(\{\omega\}) = 1/2^n$. Suppose that we are interested only in the number of heads appearing in such an experiment instead of finer details such as whether or not heads appeared on a certain flip. Observe that we can get any possible number of heads from $k = 0$ to $k = n$ by simply constructing a sequence with k many heads and the rest tails. For each $k = 0, 1, \ldots, n$, we ask what is the probability of the event E_k that there are exactly k heads and $n-k$ tails? In the same way that we argued for the binomial theorem, we see that, for each such k, there are $\binom{n}{k}$ sequences with exactly k heads and $n-k$ tails and, as we mentioned above, each sequence has probability $1/2^n$. Thus,

$$\mathbb{P}(E_k) = \binom{n}{k} \frac{1}{2^n}.$$

With this in mind, we can actually coarsen our sample space to simply account for the number of heads. In other words, consider the sample space

$$\Omega_n' = \{0, 1, 2, \ldots, n\}$$

where each $k \in \Omega_n'$ is the outcome that exactly k heads came up in n flips. We aim to construct a probability measure $\mathbb{P}'$ on Ω_n' that describes this situation by virtue of Proposition 2.3 (or Proposition 2.7). To this end, define

$$p_k = \mathbb{P}(E_k) = \binom{n}{k}\frac{1}{2^n}$$

for each $k \in \Omega_n'$; this is clearly an assignment of non-negative numbers to the outcomes in Ω_n'. To verify that this collection defines a probability measure, we must simply check that $p_0, p_1, \ldots, p_n$ sum up to 1. This is made easy by the binomial theorem: We have

$$\sum_{k=1}^{n} p_k = \sum_{k=0}^{n}\binom{n}{k}\frac{1}{2^n} = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{2}\right)^{n-k}\left(\frac{1}{2}\right)^{k} = \left(\frac{1}{2}+\frac{1}{2}\right)^{n} = 1^n = 1,$$

as required. Hence $p_0, p_1, \ldots, p_n$ defines a probability measure $\mathbb{P}'$ on this coarser sample space Ω_n' where in the probability of flipping a fair coin and seeing exactly k heads is,l

$$p_k = \mathbb{P}'(\{k\}) = \binom{n}{k}\frac{1}{2^n}.$$

Chapter 4

Conditioning and Independence

This chapter is about knowledge and its transference. We will be concerned with how knowing something affects our understanding of something else. In probabilistic terms, these ideas come by way of "conditional probability" and "independence". These are two terms that you, very likely, have good intuition about. When we formalize them and develop their theory, we will find some amazingly non-intuitive things.

4.1 Conditional Probability

If I have two events E and F in a probability space Ω equipped with probability measure $\mathbb{P}$, I could ask: If I know that the event F happened (or didn't), what does that tell me about E and its probability? If we interpret $\mathbb{P}(E)$ to be the likelihood of E, does this likelihood change provided that I know F happens or will happen? As we will see, the following definition is key to answering this question.

Definition 4.1. *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. Given two events $E, F \subseteq \Omega$ with $\mathbb{P}(F) > 0$, we define the conditional probability of E given F by*

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)}$$

Example 4.1: Simple Conditioning

Consider an experiment where we flip two coins and regard the four outcomes HH, HT, TH,and TT as equally likely. In other words, $\Omega = \{HH, HT, TH, TT\}$ and $\mathbb{P}(E) = \#(E)/4$ is the uniform measure. Consider the event F that tails came up on the second flip, i.e., $F = \{HT, TT\}$, and let's compute the conditional probabilities of various events given F:

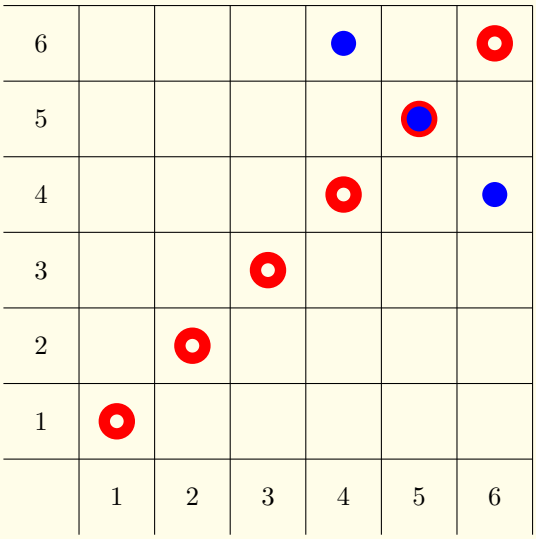
1. Consider the single outcome event that two heads came up, i.e., $E = \{HH\}$. Though the probability of the event E is $1/4$, if F happens, then E cannot happen as $EF = E \cap F = \varnothing$. Rightly so, the probability of E given F is
$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(\varnothing)}{1/2} = 0.$$

2. Consider the single outcome event $E = \{TT\}$. Though the probability of E is again $1/4$ as it occupies one quarter of the sample space, if we know that F happens, we know that the outcomes HH and TH did not happen and so it seems that the probability of E should then increase. Here, the probability of E given F is
$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(\{TT\})}{1/2} = \frac{1/4}{1/2} = \frac{1}{2}.$$
Indeed, knowing F does affect the probability of E. This is notably because E occupies half of F.

3. Finally, consider the event that tails appears on the first flip, i.e., $E = \{TH, TT\}$. In this case, $\mathbb{P}(E) = 1/2$ and
$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(\{TT\})}{1/2} = \frac{1/4}{1/2} = \frac{1}{2}.$$
In other words, the event that tails appears on the second flip doesn't tell us anything about whether or not I get tails on the first flip.

Example 4.2: Rolling Two Dice

Let's suppose that we roll two perfect dice and regard all possible outcomes as equally likely. What is the probability that the sum of the face values is 10 given that doubles are rolled?
For this example it is instructive to visualize the sample space Ω as the 6×6 array shown below.



We can consider the horizontal direction as encoding the value of the first die face and, likewise, the value of the second die is encoded in the vertical direction. Given our assumption that all events are equally likely, each ordered pair above has probability $1/36$. As can be seen in the array, the event E that the sum of the face values is 10 appears with the outcomes $(4, 6)$, $(5, 5)$ and $(6, 4)$ which is illustrated by the red dots. The event F that doubles are rolled is illustrated by the six red annulli at the points $(1, 1)$, $(2, 2)$, $(3, 3)$, $(4, 4)$, $(5, 5)$ and $(6, 6)$. By definition of conditional probability, we have
$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(5, 5)}{\mathbb{P}(\{(1, 1), (2, 2,), (3, 3), (4, 4), (5, 5), (6, 6)\})} = \frac{1}{6}$$

Let's argue that we can produce this answer in another, perhaps more intuitive, way. Since we are given the event F that doubles are rolled, our sample space (or universe) reduces from the 36 possible outcomes in the array to simply those six along the diagonal. Now, the event E that the sum of faces is a 10, in our updated sample space, can only happen provided that two fives are rolled. As each diagonal outcome is equally likely (and there are six of them) and only one of them has a face sum of 10, the probability we seek must be $1/6$.

Exercise 4.1: Throwing Three Perfect Dice

Suppose that three perfect dice are thrown and all outcomes are equally likely. If it is known that all three faces are different, use the definition of conditional probability to compute the probability that

1. At least one face is 2.
2. The sum of faces is 9

Is there a way to answer the first question without using conditional probability? If so, explain.

As discussed in Example 4.1, in essence, conditioning by F, i.e., the process of computing conditional probabilities of events given F, is the process of restricting your new sample space (or universe) to F. Along these lines, we have the following.

Proposition 4.2. *Let Ω be a sample space equipped with probability measure $\mathbb{P}$ and let F be an event with $\mathbb{P}(F) > 0$. Then the assignment*

$$E \mapsto \mathbb{P}(E|F)$$

of events $E \subseteq F$ to numbers is itself a probability measure on the "new" sample space F.

Proof. It is clear that, for each $E \subseteq F$, $\mathbb{P}(E|F) \geq 0$ and so Axiom 1 holds. Now, let $E_1, E_2, \ldots$ be a collection of disjoint events (all subsets of F). To verify Axiom 2, we must show that

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} E_k \middle| F\right) = \sum_{k=1}^{\infty} \mathbb{P}(E_k|F).$$

Since all events E_k are subsets of F, intersecting any (and all) events with F does nothing. In particular, $E_k = E_k F$ for all k and

$$\left(\bigcup_{k=1}^{\infty} E_k\right) \cap F = \bigcup_{k=1}^{\infty} E_k.$$

Using the additivity of $\mathbb{P}$, we have

$$\mathbb{P}\left(\left(\bigcup_{k=1}^{\infty} E_k\right) \cap F\right) = \mathbb{P}\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(E_k) = \sum_{k=1}^{\infty} \mathbb{P}(E_k F).$$

Therefore

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} E_k \middle| F\right) = \frac{1}{\mathbb{P}(F)} \mathbb{P}\left(\left(\bigcup_{k=1}^{\infty} E_k\right) \cap F\right) = \frac{1}{\mathbb{P}(F)} \sum_{k=1}^{\infty} \mathbb{P}(E_k F) = \sum_{k=1}^{\infty} \frac{\mathbb{P}(E_k|F)}{\mathbb{P}(F)} = \sum_{k=1}^{\infty} \mathbb{P}(E_k|F).$$

Finally, we see that $\mathbb{P}(F|F) = \mathbb{P}(F)/\mathbb{P}(F) = 1$ and so Axiom 3 i satisfied. □

Quite often, it will be the case that we will want to know the conditional probability of an event E given F but instead know the conditional probability of F given E. For example, I might want to find the probability that it will rain given that it is cloudy where I know the probability that it is cloudy given that it rains (= 1). If I know the probability of the two events E and F, it turns out that I can compute my answer. This is:

Theorem 4.3 (Bayes' Theorem 1)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. If E and F are both events with positive probability, then*

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(F)\mathbb{P}(E|F)}{\mathbb{P}(E)}.$$

Example 4.3: It's Cloudy!

Suppose that it rains 40 days of the year and it is cloudy 120 days per year. On any day on which it is cloudy, what is the probability that it will rain?

To answer this question, let's first rephrase in terms of conditional probability. We want to compute the conditional probability of $R = \{\text{It rains}\}$ given that $C = \{\text{It's cloudy}\}$. In this case, we have $\mathbb{P}(C) = 120/365$, $\mathbb{P}(R) = 40/365$ and, since it must be cloudy to rain, $\mathbb{P}(C|R) = 1$. Using the proposition, we have

$$\mathbb{P}(R|C) = \frac{\mathbb{P}(R)\mathbb{P}(C|R)}{\mathbb{P}(C)} = \frac{(40/365)(1)}{120/365} = \frac{40}{120} = \frac{1}{3}.$$

Exercise 4.2: Was it heads or tails?

A fair coin is flipped. If "heads", appears then one (perfect) die is cast coming up with a number N from 1-6 which is the die's face value. If "tails", appears, then two (perfect) dice are cast and we associate a number N (from 2-12) to the sum of their face values.

1. If "heads" appears, what's the probability that $N = 6$?
2. If "tails" appears, what's the probability that $N = 6$?
3. If $N = 6$, what's the probability that "heads" was the result of the coin flip? What is the probability that "tails" is the result? Hint: To compute $\mathbb{P}(N = 6)$ it's helpful to note that $\{\{H\}, \{T\}\}$ is a partition of Ω and so the event $\{N = 6\}$ can be expressed as the disjoint union $(\{N = 6\} \cap \{H\}) \cup (\{N = 6\} \cap \{T\})$.
4. If $1 \leq N \leq 3$, what is the probability that the coin flip resulted in "heads"?

We complete this subsection by presenting a useful formula called the multiplication rule.

Proposition 4.4. *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. Given any finite collection of events $E_1, E_2, \ldots, E_n$,*

$$\mathbb{P}(E_1E_2 \cdots E_n) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2) \cdots \mathbb{P}(E_n|E_1E_2 \cdots E_{n-1})$$

provided that $\mathbb{P}(E_1E_2 \cdots E_{n-1}) > 0$.

Remark 4.5. Given that $E_1E_2 \cdots E_{n-1} \subseteq E_1E_2 \cdots E_{n-2} \subseteq \cdots \subseteq E_1E_2 \leq E_1$, the monotonicity of probability shows that $\mathbb{P}(E_1E_2, \ldots E_k) \geq \mathbb{P}(E_1E_2 \cdots E_{n-1}) > 0$ for all $k = 1, 2, \ldots, n-1$ and thus all conditional probabilities above are defined.

Proof. I'll argue the formula for $2 \leq n \leq 3$ and invite you to verify it for larger n (Hint: Use induction). Since $\mathbb{P}(E_2|E_1) = \mathbb{P}(E_1E_2)/\mathbb{P}(E_1)$, we have

$$\mathbb{P}(E_1E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)$$

which is the formula for $n = 2$. Also, because $\mathbb{P}(E_3|E_1E_2) = \mathbb{P}(E_1E_2E_3)/\mathbb{P}(E_1E_2)$, we have, by virtue of the $n = 2$ result,

$$\mathbb{P}(E_1E_2E_3) = \mathbb{P}(E_1E_2)\mathbb{P}(E_3|E_1E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2).$$

□

here

Example 4.4: Rein over your own pile!

Let's take an ordinary deck of cars, shuffle it, and divide it into 4 piles of 13. What is the probability that all kings are in separate piles?

I hope you agree that this is an interesting question that, a priori, seems to have little to do with conditional probability. In fact, this is somewhat a nightmare of a counting problem if done directly. Thankfully, the multiplication formula comes to the rescue. Let K_1, K_2, K_3, and K_4 denote the four kings. Let E_1 denote the event that K_1 is in one of the piles; note, this has to be true and so $\mathbb{P}(E_1) = 1$. Let E_2 denote the event that K_1 and K_2 are in separate piles. Continuing, E_3 denotes the event that K_1, K_2 and K_3 are in separate piles and, finally, E_4 denotes the event that all kings are in separate piles.

Let's note first that the events E_1, E_2, E_3, and E_4 are nested and so

$$E_2 = E_1E_2, \quad E_3 = E_1E_2E_3, \text{ and } E_4 = E_1E_2E_3E_4.$$

With this observation, the multiplication rule gives

$$\begin{aligned}\mathbb{P}(E_4) = \mathbb{P}(E_1E_2E_3E_4) &= \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2)\mathbb{P}(E_4|E_1E_2E_3)\\ &= \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_2)\mathbb{P}(E_4|E_3)\\ &= \mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_2)\mathbb{P}(E_4|E_3)\end{aligned}$$

where we have used the fact that $\mathbb{P}(E_1) = 1$. To compute the conditional probability $\mathbb{P}(E_2|E_1)$, we can think in terms of open slots in which cars can be placed. One slot is taken by K_1 and so, K_2 can be placed into the 51 remaining slots. Observe that there are $39 = 51 - 12$ slots in which K_2 can be placed into a separate pile from K_1. Hence, the probability that K_1 and K_2 are placed into separate piles is precisely the probability $\mathbb{P}(E_2|E_1) = 39/51$. Now, the conditional probability $\mathbb{P}(E_3|E_2)$ is the probability of the event that K_3 is placed into a pile separate of K_1 and K_2 which occupy two different piles themselves. So, if we think in terms of slots, there are 50 possible slots for K_3, but only $26 = 13 + 13 = 50 - 12 - 12$ are slots within piles that do not contain K_1 or K_2. Hence $\mathbb{P}(E_3|E_2) = 26/50$. Finally, by a similar argument (which you should make yourself), $\mathbb{P}(E_4|E_3) = 13/49$. All together,

$$\mathbb{P}(E_4) = \left(\frac{39}{51}\right)\left(\frac{26}{50}\right)\left(\frac{13}{49}\right) = \frac{2197}{20825} \approx 0.105$$

4.1.1 The Law of Total Probability

Often times, one can understand the probabilities of events in a sample space by knowing their probabilities conditioned on several different things. Our first big result, the law of total probability, will tell us how to use these conditional probabilities to compute the (unconditional) probability of the event. With it, we shall immediately obtain a second (and very useful) form of Bayes' theorem.

Theorem 4.6 (The Law of Total Probability)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$.*

1. *If E and F are events, then*
$$\mathbb{P}(E) = \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c).$$
2. *More generally, if E is an event and $F_1, F_2, \ldots$ is a collection of events forming a partition of Ω, then*
$$\mathbb{P}(E) = \sum_{k=1}^{\infty} \mathbb{P}(E|F_k)\mathbb{P}(F_k).$$

Remark 4.7. Technically, $\mathbb{P}(E|F)$ isn't defined when $\mathbb{P}(F) = 0$. However, with the interpretation that $\mathbb{P}(E|F)\mathbb{P}(F) = 0$ whenever $\mathbb{P}(F) = 0$, the above formulas hold in general.

Example 4.5: Car Insurance

Car insurance companies treat teenage drivers as being more prone to accidents than their non-teenage counterparts. A given company's statistics show that a teenager will have an accident at some time in a six-month period with a probability of 2/5. By contrast, the company estimates that a non-teenage driver has a probability of 1/7 of getting into an accident in the same six-month period. If 1/10 of the company's insurance holders are teens, what is the probability that a randomly selected driver (teenage or not) will have an accident in a six-month period?
To answer this question, let E be the event that the randomly selected driver has an accident and F be the event that the driver is a teenager. In our set-up,

$$\mathbb{P}(E|F) = \mathbb{P}(\text{The driver will have an accident given that they are a teen}) = \frac{2}{5}$$

and

$$\mathbb{P}(E|F^c) = \mathbb{P}(\text{The driver will have an accident given that they are not a teen}) = \frac{1}{7}.$$

Since the proportion of teen drivers to drivers overall is 1/10, we have $\mathbb{P}(F) = 1/10$, $\mathbb{P}(F^c) = 1-\mathbb{P}(F) = 9/10$ and so, by the law of total probability,

$$\mathbb{P}(E) = \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c) = \left(\frac{2}{5}\right)\left(\frac{1}{10}\right) + \left(\frac{1}{7}\right)\left(\frac{9}{10}\right) = \frac{59}{350} \approx 0.17$$

Example 4.6: Mike Goes to the Casino

Michael likes to go to the casino. When he does, he likes to visit the blackjack tables, the poker tables and the slot machines with equal proportions (1/3). When he plays blackjack, he wins money with probability 2/5, when he plays poker, he wins with probability 1/5, and, when he plays the slots, he wins with probability 1/20. In general, what is the probability that Michael wins money when he visits the casino?

It is clear that the events F_1, F_2, and F_3 that Michael visits the blackjack tables, poker tables, and slot machines, respectively, are mutually exclusive and collectively exhaustive (thus forming a partition). If E denotes the event that Michael wins money at the casino, we have

$$\mathbb{P}(E|F_1) = \mathbb{P}(\text{Michael wins money given that he plays blackjack}) = \frac{2}{5}$$

and, similarly, we see that $\mathbb{P}(E|F_2) = 1/5$ and $\mathbb{P}(E|F_3) = 1/20$. Further, given that he plays these games in equal proportion, we have $\mathbb{P}(F_1) = \mathbb{P}(F_2) = \mathbb{P}(F_3) = 1/3$. Thus, by the law of total probability,

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E|F_1)\mathbb{P}(F_1) + \mathbb{P}(E|F_2)\mathbb{P}(F_2) + \mathbb{P}(E|F_3)\mathbb{P}(F_3)\\ &= \left(\frac{2}{5}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{5}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{20}\right)\left(\frac{1}{3}\right)\\ &= \frac{13}{60}\end{aligned}$$

Exercise 4.3: Get out the vote!

In a certain voting district, 10% identify as independent, 52% as liberal and 38% as conservative. Despite their identification, these voters are not equally reliable in showing up to vote. In a given election, it is known that 34% of independents voted, 51% of liberals voted, and 79% of conservatives voted.

1. If you select a random person from the community (who is registered to vote), what is the probability that they voted in the election?
2. If the person you did select voted, what is the probability that they are conservative? What is the probability that they are liberal? What is the probability that they are independent?
3. If each party runs one candidate for office and voters vote along their party lines, which candidate wins?

Proof of Theorem 4.6. We note that, for any event F, F and F^c together form a partition of Ω. With this, we see that the first item is a special case of the first [1]. Thus, we shall prove the second item. Given that $F_1, F_2, \ldots$ forms a partition of Ω, we are able to write

$$E = \bigcup_k EF_k$$

and therefore

$$\mathbb{P}(E) = \sum_{k=1}^{\infty} \mathbb{P}(EF_k)$$

by virtue of Exercise 2.5. Since $\mathbb{P}(EF_k) = \mathbb{P}(E|F_k)\mathbb{P}(F_k)$ for each k, we have

$$\mathbb{P}(E) = \sum_{k=1}^{\infty} \mathbb{P}(E|F_k)\mathbb{P}(F_k),$$

as desired. □

Combining the results of Theorem 4.3 (for $F = F_n$) and Theorem 4.6, we immediately obtain the second form of Bayes' theorem.

Theorem 4.8 (Bayes' Theorem II)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$.*

1. *If E and F are events with positive probability, then*

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F)\mathbb{P}(F)}{\mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c)}$$

2. *More generally, E is an event with positive probability and $F_1, F_2, \ldots$ is a collection of events forming a partition of Ω, then, for each n,*

$$\mathbb{P}(F_n|E) = \frac{\mathbb{P}(E|F_n)\mathbb{P}(F_n)}{\sum_k \mathbb{P}(E|F_k)\mathbb{P}(F_k)}$$

Example 4.7: False Positives

Suppose that a blood test is 96% effective at detecting a disease when it is present. However, for every 1 in 50 people who do not have the disease, the test indicates that they do; this is called a *false positive*. If 2% of the population actually does have the disease, what is the probability that a person has the disease

[1] If you're worried about $\{F, F^c\}$ being a finite partition and not a countably infinite one, feel free to adjoin an infinite number of empty sets together with F and F^c to form a countable partition – the result is the same.

given that they test positive?

Let's denote by P the event that a person tests positive and D the event that they have the disease. Our aim is to compute $\mathbb{P}(D|P)$. Of course, none of the information given to us in this form. We do know, however, someone's likelihood of testing positive given that they have the disease, $\mathbb{P}(P|D)$, and the likelihood that they test positive given that they do not have the disease, $\mathbb{P}(P|D^c)$. Thus, by the above version of Bayes' theorem, we have

$$\begin{aligned}\mathbb{P}(D|P) &= \frac{\mathbb{P}(P|D)\mathbb{P}(D)}{\mathbb{P}(P|D)\mathbb{P}(D)+\mathbb{P}(P|D^c)\mathbb{P}(D^c)}\\ &= \frac{(96/100)(2/100)}{(96/100)(2/100)+(1/50)(98/100)}\\ &= \frac{48}{97}\approx 0.495.\end{aligned}$$

That's a surprisingly low number! Wouldn't you want a better test?

Example 4.8: A Card Game

Suppose that we have three cards that are identical in form, except that both sides of the first card are colored red, both sides of the second card are colored black, and the last card is black on one side and red on the other. If we select a card at random, place it down on the table and the face showing is red, what is the probability that the other side is black?

Let's denote the events by C_{rr}, C_{bb} and C_{rb}, the events that our chosen card was the double red, double black and red-black card, respectively. Let's also denote the event F_r as the event that the face placed up on the table is red. Our aim is to compute $\mathbb{P}(C_{rb}|F_r)$. Given that C_{rr}, C_{bb}, and C_{rb} form a partition of Ω, Bayes' theorem guarantees that

$$\begin{aligned}\mathbb{P}(C_{rb}|F_r) &= \frac{\mathbb{P}(F_r|C_{rb})\mathbb{P}(C_{rb})}{\mathbb{P}(F_r|C_{rb})\mathbb{P}(C_{rb})+\mathbb{P}(F_r|C_{rr})\mathbb{P}(C_{rr})+\mathbb{P}(F_r|C_{bb})\mathbb{P}(C_{bb})}\\ &= \frac{\mathbb{P}(F_r|C_{rb})\left(\frac{1}{3}\right)}{\mathbb{P}(F_r|C_{rb})\left(\frac{1}{3}\right)+\mathbb{P}(F_r|C_{rr})\left(\frac{1}{3}\right)+\mathbb{P}(F_r|C_{bb})\left(\frac{1}{3}\right)}\\ &= \frac{\mathbb{P}(F_r|C_{rb})}{\mathbb{P}(F_r|C_{rb})+\mathbb{P}(F_r|C_{rr})+\mathbb{P}(F_r|C_{bb})}\end{aligned}$$

where we have noted that $\mathbb{P}(C_{rr})+\mathbb{P}(C_{bb})+\mathbb{P}(C_{rb})=1/3$, i.e., that each card had an equal chance of being selected. Now, if the red-black card was selected, the conditional probability $\mathbb{P}(F_r|C_{rb})=1/2$ since each side is equally likely to come up and only one of those sides is red. However, since the C_{rr} card has two red sides, $\mathbb{P}(F_r|C_{rr})=1$ because, with one hundred percent probability, a red face shows. By a similar argument, $\mathbb{P}(F_r|C_{bb})=0$. Thus,

$$\mathbb{P}(C_{rb}|F_r)=\frac{\left(\frac{1}{2}\right)}{\left(\frac{1}{2}\right)+1+0}=\frac{1}{3}.$$

Had we instead phrased this problem by saying that there are six possible faces that can be placed "up" on the table, each of which is equally likely, then the conditional probability that the red-black card was chosen given that red came up is equivalent to asking which proportion of the total red faces belong to the red-black card. This is, of course, 1/3.

Exercise 4.4: Plane Crash!

A plane goes missing and it is assumed that the plane went down in one of three regions each of which is equally likely. Let $\alpha_1 = 1 - \beta_1$ denote the probability that the plane is found upon searching the first region given that it actually went down in the first region. More generally, for $k = 1, 2, 3$, $\alpha_k = 1 - \beta_k$ represents the probability that the plane is found in region k given that it actually went down there and we assume that $0 < \alpha_k < 1$ for $k = 1, 2, 3$. The numbers α_1, α_2, and α_3 are typically called *search probabilities* and depend on the typography of the region in question. If it is known that the search for the plane in the first region is unsuccessful, what is the probability that the plane went down in that region (despite that it wasn't found there)?

here

4.2 Independence

In the previous section, we introduced the conditional probability of an event E given (or conditioned on) another event F and argued that this conditional probability $\mathbb{P}(E|F)$ describes the likelihood of the event E if we "know" that the event F is to happen or has happened. In this study, we've seen a few examples where knowing F didn't actually change the probability of E. Let's revisit Example 4.1 to illustrate this point.

Example 4.9: Two Fair Coins

As in Example 4.1, we flip two fair coins and regard the outcomes in

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

are equally likely. We consider the event F that the tails appears on the first coin flip and the event E that tails appears on the second coin flip. As we saw,

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(\{TT\})}{\mathbb{P}(\{TH, TT\})} = \frac{1/4}{1/2} = \frac{1}{2}.$$

And this coincides with the unconditional probability

$$\mathbb{P}(E) = \mathbb{P}(\{HT, TT\}) = \frac{1}{2}.$$

Hence,

$$\mathbb{P}(E|F) = \mathbb{P}(E)$$

By a similar computation, it is easy to see that the same equation holds if E is replaced by the event that heads is observed on the second flip. Overall, we can interpret these observations in the following way: "Knowing" that tails appeared on the first flip didn't actually affect the outcome of the second flip. In this way, we might think of these flips as being done freely or, perhaps, independently.

In general, on a sample space Ω equipped with probability measure $\mathbb{P}$, consider two events E and F such that

$$\mathbb{P}(E|F) = \mathbb{P}(E).$$

In this case, we have

$$\mathbb{P}(EF) = \mathbb{P}(E|F)\mathbb{P}(F) = \mathbb{P}(E)\mathbb{P}(F).$$

Conversely, if we assume that $\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$ with $\mathbb{P}(F) > 0$, it is easy to see that

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(E)\mathbb{P}(F)}{\mathbb{P}(F)} = \mathbb{P}(E).$$

With these observations in mind, let's give this situation a name.

Definition 4.9. *Let Ω be a sample space equipped with a probability measure $\mathbb{P}$.*

1. *We say that events E and F are independent if*

$$\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F).$$

Here, E and F are also said to be pairwise independent and we sometimes write $E \perp F$.

2. *We say that a collection of three events E, F, and G are independent if*

$$\mathbb{P}(EFG) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G)$$

and $\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$, $\mathbb{P}(EG) = \mathbb{P}(E)\mathbb{P}(G)$, and $\mathbb{P}(FG) = \mathbb{P}(F)\mathbb{P}(G)$.

3. *More generally, a finite collection of events $E_1, E_2, \ldots E_n$ is said to be independent if, for every subcollection $E_{j_1}, E_{j_2}, \ldots, E_{j_k}$, we have*

$$\mathbb{P}(E_{j_1}E_{j_2}\cdots E_{j_k}) = \mathbb{P}(E_{j_1})\mathbb{P}(E_{j_2})\cdots\mathbb{P}(E_{j_k}).$$

Remark 4.10. New students to probability tend to confuse independence with the notion of being disjoint. These are entirely different concepts.

Remark 4.11. If you are familiar with product notation, the final displayed equation can be written equivalently by

$$\mathbb{P}\left(\bigcap_{l=1}^{k} E_{j_l}\right) = \prod_{l=1}^{k} \mathbb{P}(E_{j_l}).$$

Remark 4.12. In view of the discussion preceding the definition, you might wonder: For two events E and F, why don't we just say that E and F are independent when $\mathbb{P}(E|F) = \mathbb{P}(E)$ instead of the related condition $\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$? The answer is that even writing down $\mathbb{P}(E|F)$ assumes (implicitly) that $\mathbb{P}(F) > 0$ whereas the condition $\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$ requires no such assumption and, in fact, we will still find it meaningful in times when $\mathbb{P}(F)$ or $\mathbb{P}(E)$ are zero.

Example 4.10: A three part experiment

Let's consider an experiment where we roll one perfect die, flip one fair coin, and choose one card from a well-shuffled deck of 52. In this situation, we can represent our experiment with the sample space

$$\Omega = \{(\omega_1, \omega_2, \omega_3) : \omega_1 \in \{1, 2, \ldots, 6\},\ \omega_2 \in \{H, T\},\ \omega_3 \in \{1, 2, \ldots, 52\}\}.$$

If the flip, roll, and card selection are done freely, it is reasonable to expect that the $6 \times 2 \times 52 = 624$ outcomes in Ω are equally likely. With our definition of independence, we should start to think of "being done freely" as being independent. To illustrate this, let's consider the events:

$$\begin{aligned} F &= \{\text{die roll is Five}\} \\ H &= \{\text{coin flip is Heads}\} \\ K &= \{\text{card is a King}\} \end{aligned}$$

By the fundamental rule of counting, it is easy to see that

$$\mathbb{P}(F) = \frac{\#(F)}{\#(\Omega)} = \frac{1 \times 2 \times 52}{6 \times 2 \times 52} = \frac{1}{6},$$

$$\mathbb{P}(H) = \frac{\#(H)}{\#(\Omega)} = \frac{6 \times 1 \times 52}{6 \times 2 \times 52} = \frac{1}{2},$$

and

$$\mathbb{P}(K) = \frac{\#(K)}{\#(\Omega)} = \frac{6 \times 2 \times 4}{6 \times 2 \times 52} = \frac{1}{13}.$$

Also, by the fundamental rule of counting,

$$\mathbb{P}(FH) = \mathbb{P}(\text{roll is Five and flip is Heads}) = \frac{1 \times 1 \times 52}{6 \times 2 \times 52} = \frac{1}{12}.$$

Similarly,

$$\mathbb{P}(FK) = \mathbb{P}(\text{roll is Five and card is King}) = \frac{1 \times 2 \times 4}{6 \times 2 \times 52} = \frac{1}{78},$$

$$\mathbb{P}(HK) = \mathbb{P}(\text{coin is Heads and card is King}) = \frac{6 \times 1 \times 4}{6 \times 2 \times 52} = \frac{1}{26}.$$

With this, we observe that

$$\mathbb{P}(FH) = \frac{1}{12} = \frac{1}{6}\frac{1}{2} = \mathbb{P}(F)\mathbb{P}(H),$$

$$\mathbb{P}(FK) = \frac{1}{78} = \frac{1}{6}\frac{1}{13} = \mathbb{P}(F)\mathbb{P}(K),$$

and

$$\mathbb{P}(HK) = \frac{1}{26} = \frac{1}{2}\frac{1}{13} = \mathbb{P}(H)\mathbb{P}(K)$$

so that $F \perp H$, $F \perp K$, and $H \perp K$. Finally, we observe that, by the fundamental rule of counting,

$$\mathbb{P}(FHK) = \mathbb{P}(\text{roll is Five, flip is Heads, and card is King}) = \frac{1 \times 1 \times 4}{6 \times 2 \times 52} = \frac{1}{156}.$$

We therefore see that

$$\mathbb{P}(FHK) = \frac{1}{156} = \frac{1}{6}\frac{1}{2}\frac{1}{13} = \mathbb{P}(F)\mathbb{P}(H)\mathbb{P}(K)$$

and so we may conclude that the events F, H, and K are independent. You are encouraged to verify that any triple of events in this experiment, where the first depends only on the die roll, the second on the coin flip and the third on the card selection, will necessarily be independent.

As the following example shows, it is possible for three events E, F, and G to have $E \perp F$, $F \perp G$ and $E \perp G$ but the collection not be independent[2].

Exercise 4.5: More Dice

Consider the experiment of throwing two dice. Let E be the event that the sum of the outcomes is 7, F be the outcome that the first die landed on 3 and G be the event that the second die landed on 4.

1. Show that E and F are independent.
2. Show that F and G are independent.
3. Show that E and G are independent.
4. Show that (together) E, F, and G is not an independent collection

[2]Such is an example where events are pairwise independent but not all together independent.

Exercise 4.6: How many independent events can there be?

In this exercise, you will see that, in a finite sample space equipped with the uniform probability, there can only be so many independent events.

1. Let Ω be a sample space with n outcomes, i.e., $\#(\Omega) = n$, and let $\mathbb{P}$ be the uniform probability on Ω. Given $0 < p < 1$, let $E_1, E_2, \ldots, E_m$ be a collection of independent events for which $0 < P(E_k) \leq p$ for all $k = 1, 2, \ldots, m$. Show that
$$\frac{1}{n} \leq p^m.$$

2. Rearrange the above inequality to see that
$$m \leq \frac{\log(n)}{\log(1/p)}.$$
Does this inequality tell you anything about the events E, F and G in the preceding exercise? Please explain.

We conclude this subsection by stating some basic properties of independent events.

Proposition 4.13. *Let Ω be a sample space equipped with a probability measure $\mathbb{P}$ and let E, F, and G be events.*

1. *Independence of two events is commutative, i.e., $E \perp F = F \perp E$.*
2. *E and F are independent if and only if E and F^c are independent.*
3. *If E, F and G are independent, then E, F, and G^c are independent.*
4. *If E, F and G are independent, then E and $F \cup G$ are independent.*

Thinking in terms of conditional probability as a way to measure the transference of knowledge, in the case that $\mathbb{P}(F)$ and $\mathbb{P}(F^c) > 0$, the second property shows that
$$\mathbb{P}(E|F) = \mathbb{P}(E) \qquad \text{if and only if} \qquad \mathbb{P}(E|F^c) = \mathbb{P}(E).$$
In other words, if you learn nothing about E by knowing F, you cannot learn anything about E by knowing F's complement and vice versa. The fourth property has a similar interesting interpretation. It says that if E, F, and G are independent and $\mathbb{P}(F \cup G) > 0$, then
$$\mathbb{P}(E|F \cup G) = \mathbb{P}(E).$$
In other words, if we cannot learn anything about E from F or G, we cannot learn anything about E by combining our knowledge of F and G.

Proof. Here, we shall prove the first two statements and leave the second two as an exercise.

1. Since $EF = FE$ and multiplication of real numbers is a commutative operation, we have
$$\mathbb{P}(FE) = \mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F) = \mathbb{P}(F)\mathbb{P}(E)$$
and so $E \perp F$ if and only if $F \perp E$.

2. For two events E and F, E can be expressed in the disjoint union $E = EF \cup EF^c$ and so it follows that
$$\mathbb{P}(EF^c) = \mathbb{P}(E) - \mathbb{P}(EF).$$
Thus, if $E \perp F$, we have
$$\mathbb{P}(EF^c) = \mathbb{P}(E) - \mathbb{P}(E)\mathbb{P}(F) = \mathbb{P}(E)(1 - \mathbb{P}(F)) = \mathbb{P}(E)\mathbb{P}(F^c)$$
whence $E \perp F^c$. The converse follows easily by applying the above argument to E and F^c and noting that $(F^c)^c = F$.

□

Exercise 4.7: Independence Properties

1. Prove the third assertion of the proposition above. Argue why it shows that taking any collection of E, F, G and their complements are also independent.
2. Prove the fourth assertion of the proposition above.
3. Give an example of the fourth assertion. In other words, find an experiment with indpeendent events E, F, and G and verify that $E \perp F \cup G$. Hint: You can use events in Example 4.12 if you'd like.

Exercise 4.8: More Independence Properties

Let Ω be a sample space and $\mathbb{P}$ be a Probability measure on Ω. Given an event $E \subseteq \Omega$, we say that E occurs *almost surely* if $\mathbb{P}(E) = 1$.

1. Given events E and F, show that the only way for them to be independent and disjoint is for one of them to have probability zero.
2. Given events E and F, show that E and F are independent whenever one of them occurs almost surely.
3. What events are independent of themselves?

4.2.1 Independent Trials

With the notion of independence at our fingertips, we are ready to construct long-term experiments made by repeating a single "base" experiment over and over again where each sub experiment if performed independently of all those before it. For example, we could consider a single flip of a coin (the base experiment) done over and over again where each flip is done independently of those before it. This type of construction is known as *independent trials* and we will find it central our studies throughout these notes. Before talking about probabilities and independence, let's first talk about sample spaces. Let Ω_0 be a sample space, called a **base sample space**, and, for some N, consider the sample space

$$\Omega_N = \{\omega = (\omega_1, \omega_2, \ldots, \omega_N) : \omega_k \in \Omega_0 \text{ for all } k = 1, 2, \ldots, N\}$$

which is precisely the sample space by which we repeat the experiment Ω_0 N times. Here, we are thinking about each outcome of Ω_N as an N-tuple of outcomes $\omega = (\omega_1, \omega_2, \ldots, \omega_N)$ where $\omega_1 \in \Omega_0$ represents the outcome of the first experiment, $\omega_2 \in \Omega_0$ represents the outcome of the second experiment and so on. For example, if $\Omega_0 = \{H, T\}$ represents the sample space of a single coin flip, the outcome $\omega = (H, T, H, H, T) \in \Omega_5$ represents a sequence of five coin flips where heads appeared on the first, third and fourth flips and tails appeared on the second and last flips.

For a general base sample space Ω_0, the new sample space Ω_N is known as the N-fold Cartesian product of Ω_0 (with itself) and often written[3]

$$\Omega_N = \Omega_0^N = \underbrace{\Omega_0 \times \Omega_0 \times \cdots \Omega_0}_{N}.$$

With this construction, each individual multiplicand Ω_0 is called a **trial** and, in particular, the kth multiplicand Ω_0 is called k**th trial** and corresponds to the kth coordinate ω_k in any outcome $\omega = (\omega_1, \omega_2, \ldots \omega_{k-1}, \omega_k, \omega_{k+1}, \ldots, \omega_N) \in \Omega_N$. We shall soon discuss the case in which $N = \infty$, i.e., where we repeat the same base experiment over and over *ad infinitum*, for now let's just focus on a finite N and discuss how to equip Ω_N with a probability measure in which the trials are independent.

[3]This is precisely the way that the Cartesian plane $\mathbb{R}^2$ is gotten by forming ordered pairs (x, y) from the base $\mathbb{R}$ so that $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$.

In the special case that Ω_0 is finite with $\#(\Omega_0) = m$, (perhaps) the easiest way to construct a probability measure on Ω_N in which the trials are independent is to simply use the uniform probability measure $\mathbb{P} = \mathbb{P}_u$. In this case, the fundamental rule of counting shows that

$$\mathbb{P}(E) = \frac{\#(E)}{m^N}.$$

for any event $E \in \Omega_N$. With this measure, we can easily show that the trials are independent, i.e., anything that happens in one trial is independent of all others. Let's give this argument for a simple pair of events, the first depending only on the jth trial and the second depending only on the kth trial.

Consider consider any two events F_j and F_k in the base sample space Ω_0 (i.e., $F_j, F_k \subseteq \Omega_0$) and define

$$\begin{aligned} E_j &= \Omega_0 \times \Omega_0 \times \cdots \times F_j \times \cdots \times \Omega_n \\ &= \{\omega = (\omega_1, \omega_2, \ldots, \omega_j, \ldots, \omega_N) : \omega_j \in F_j, \text{ and } \omega_l \in \Omega_0 \text{ for all } l \neq j\}. \end{aligned}$$

and, similarly,

$$\begin{aligned} E_k &= \Omega_0 \times \Omega_0 \times \cdots \times F_k \times \cdots \times \Omega_n \\ &= \{\omega = (\omega_1, \omega_2, \ldots, \omega_k, \ldots, \omega_N) : \omega_k \in F_k, \text{ and } \omega_l \in \Omega_0 \text{ for all } l \neq k\}. \end{aligned}$$

where we ask that $j \neq k$. Though this is somewhat notationally heavy, E_j (resp. E_k) is the event that says F_j (resp. F_k) must happen in the jth (resp. kth) trial but anything can happen in all other trials. By the fundamental principle of counting, we have that

$$\mathbb{P}(E_j) = \frac{m \times m \times m \cdots \times \#(F_j) \times m \times m}{m^N} = \frac{\#(F_j)}{m} = \frac{\#(F_j)}{\#(\Omega_0)}$$

and similarly

$$\mathbb{P}(E_k) = \frac{\#(F_k)}{m} = \frac{\#(F_k)}{\#(\Omega_0)}.$$

Since $j \neq k$, the intersection of E_k and E_j is simply gotten by restricting both the jth trial to F_j and the kth trial to F_k so that (if $j < k$),

$$\begin{aligned} E_j E_k &= \Omega_0 \times \Omega_0 \times \cdot \times \Omega_0 \times F_j \times \Omega_0 \cdots \Omega_0 \times F_k \times \Omega_0 \cdots \Omega_0 \\ &= \{\omega = (\omega_1, \omega_2, \ldots, \omega_N) : \omega_l \in F_k \text{ for } l = j, k \text{ and } \omega_l \in \Omega_0 \text{ whenever } l \neq j, k\}. \end{aligned}$$

Again, by the fundamental rule of counting, we find that

$$\mathbb{P}(E_j E_k) = \frac{\#(F_j)\#(F_k)}{m^2} = \mathbb{P}(E_j)\mathbb{P}(E_k)$$

and so E_j and E_k are independent. Looking back through the above argument, you should see the uniform measure on the base floating around! In fact, if we denote by $\mathbb{P}_0$ the uniform measure on the base Ω_0, the above argument shows that $\mathbb{P}(E_j) = \mathbb{P}_0(F_j)$, $\mathbb{P}(E_k) = \mathbb{P}_0(F_k)$ and

$$\mathbb{P}(E_j E_k) = \mathbb{P}(E_j)\mathbb{P}(E_k) = \mathbb{P}_0(F_j)\mathbb{P}_0(F_k)$$

This observation turns out to be key to showing that all distinct trials are independent. Consider a collection of events $F_1, F_2, \ldots, F_N$ in the base space Ω_0 (any of them can be the full sample space Ω_0 or simply a subevent of Ω_0) and define the so-called rectangle

$$F_1 \times F_2 \times \cdots \times F_N = \{\omega = (\omega_1, \omega_2, \ldots, \omega_N) : \omega_l \in F_l \text{ for } l = 1, 2, \ldots, N\} \subseteq \Omega_N$$

By pushing our argument above, one can show (using the fundamental principle of counting) that

$$\mathbb{P}(F_1 \times F_2 \times \cdots \times F_N) = \mathbb{P}_0(F_1)\mathbb{P}_0(F_2) \cdots \mathbb{P}_0(F_N) \tag{4.1}$$

With this observation, it is easy to show that all trials are independent (with the uniform measure).

Exercise 4.9: Three coin flips

Let $\Omega_0 = \{H, T\}$ be a single experiment where a coin is flipped. Using it as the base, lets form the experiment of flipping the coin three times with sample space

$$\Omega_3 = \Omega_0^3 = \{(\omega_1, \omega_2, \omega_3) : \omega_k \in \Omega_0 \text{ for } k = 1, 2, 3\}.$$

Now, let F_1, F_2, and F_3 be any events in Ω_0 and form the events:

$$\begin{aligned} E_1 &= F_1 \times \Omega_0 \times \Omega_0 = \{(\omega_1, \omega_2, \omega_3) : \omega_1 \in F_1, \omega_2, \omega_3 \in \Omega_0\} \\ E_2 &= \Omega_0 \times F_2 \times \Omega_0 = \{(\omega_1, \omega_2, \omega_3) : \omega_2 \in F_2, \omega_1, \omega_3 \in \Omega_0\} \\ E_3 &= \Omega_0 \times \Omega_0 \times F_3 = \{(\omega_1, \omega_2, \omega_3) : \omega_3 \in F_3, \omega_1, \omega_2 \in \Omega_0\} \end{aligned}$$

Either arguing directly or making use of (4.1), show that the events E_1, E_2, and E_3 are independent. Would you agree that this shows that anything that happens in one flip of the coin is independent of all other flips?

Though we won't prove this in general, one can extend the argument used in the above exercise to show that, for a general finite sample space Ω_0 and finite N, the uniform measure on Ω_N satisfies (4.1) and, with just this property alone, one can prove that all N trials are independent. We are naturally led to two questions:

Q1 What if the base measure $\mathbb{P}_0$ is not uniform? For example, what if we flip a biased coin? In this case, can we will find a probability measure on Ω_N in which all N trials are independent?

Q2 What if we want to perform trials forever? In other words, starting with a base sample space Ω_0 consider the sample space of infinite sequences

$$\Omega_\infty = \{\omega = \omega_1, \omega_2, \omega_2 \cdots : \text{ where } \omega_k \in \Omega_0 \text{ for } k = 1, 2, \dots\}$$

representing infinite trials of the base experiment Ω_0. Clearly, there is no uniform probability measure on Ω_∞ and so our argument above falls to pieces. Can we still form a probability measure on Ω_∞ in which all trials are independent?

As it turns out, both questions can be answered in the affirmative and the key to the solution is to ask that (and appropriate generalization of) (4.1) holds.

Theorem 4.14. *Let Ω_0 be a base sample space equipped with a probability measure $\mathbb{P}_0$. There is a unique probability measure $\mathbb{P}$ on Ω_N (and we allow N to be either finite or ∞ as discussed above) that makes all trials independent. In the case that N is finite, $\mathbb{P}$ satisfies*

$$\mathbb{P}(F_1 \times F_2 \times \cdots F_N) = \mathbb{P}_0(F_1)\mathbb{P}_0(F_2) \cdots \mathbb{P}_0(F_N)$$

for any collection of events $F_1, F_2, \dots, F_N$. In the case that N is infinite, $\mathbb{P}$ satisfies

$$\mathbb{P}(E) = \prod_{k=1}^{\infty} \mathbb{P}_0(F_k)$$

whenever E is an infinite rectangle of the form

$$E = \{\omega = \omega_1, \omega_2, \cdots : \omega_k \in F_k \text{ for all } k = 1, 2, \dots\}$$

where $F_k \subseteq \Omega_0$ for all $k = 1, 2, \dots$.

Giving a proof of the above theorem is beyond the scope of these notes. In fact, the result is often taken for granted in many standard graduate probability textbooks – it is either just simply assumed to be true and not stated or

stated and not proved. In my opinion, this is quite unfortunate because the result is beautiful and powerful and is used constantly in both statistics and probability. The theorem is credited to the Danish mathematician, Børge Jessen, who proved it in the 1930s; a nice proof and history can be found in [10, Theorem 3, pp. 163].

With the help of the construction above, we are ready to introduce the most important example of infinite trials, the so-called Bernoulli trials.

Example 4.11: Bernoulli Trials

Let $0 \leq p \leq 1$ and consider a single biased coin flip $\Omega_0 = \{H, T\}$ with probability $\mathbb{P}_0$ given by

$$\mathbb{P}_0(\omega) = \begin{cases} p & \omega = H \\ q = 1 - p & \omega = T. \end{cases}$$

With this base trial Ω_0, we form the infinite sequence of trials

$$\Omega_\infty = \{\omega = \omega_1, \omega_2 \cdots : \omega_k \in \Omega_0 \text{ for } k = 1, 2, \dots\}$$

and equip Ω_∞ with the probability measure $\mathbb{P}$ given in Theorem 4.14 which ensures that, though each flip is biased landing on heads with probability p and tails with probability q, all flips are done independently. These are the **Bernoulli trials** named after the 17th-century Swiss mathematician, Jacob Bernoulli. With $\mathbb{P}$ we can answer just about any question you can ask about these repeated independent flips of a coin. For example, consider the event E_n that the coin landed on tails for the first $n-1$ flips and then heads on the nth flip. We can recognize E_n as the rectangular event

$$E_n = \{\omega = \omega_1\omega_2 \cdots : \omega_k \in F_k \text{ for all } k = 1, 2, \dots\}$$

where the "basic events" F_k are given by

$$F_k = \begin{cases} \{T\} & k = 1, 2, \dots, n-1 \\ \{H\} & k = n \\ \Omega_0 = \{H, T\} & k = n+1, n+2, \dots \end{cases}$$

Since $\mathbb{P}_0(F_k) = \mathbb{P}_0(\Omega_0) = 1$ for all $k > n$, the infinite product in Theorem 4.14 reduces to simply

$$\mathbb{P}(E_n) = \prod_{k=1}^{n} \mathbb{P}(F_k) = \underbrace{q \cdot q \cdot q \cdots q}_{n-1} \cdot p = q^{n-1}p.$$

Example 4.12: Gambler's Ruin

Let's suppose that Mike walks into a casino with an initial fortune I (in dollars) and plays a game where he gambles against the casino by performing Bernoulli trials (e.g., flipping a coin) where the probability of him winning each trial is p and the probability of losing each trial is $q = p - 1$. For any trial, if Mike wins he is given one dollar and if he loses he pays one dollar. The game ends when Mike either has N dollars (where he wins) and loses when he has 0 dollars. Here, N is the total possible money the the casino is willing to put up for this game. So, if Mike walks in with I dollars, then the casino puts up $N - I$ dollars so someone, either Mike or the casino, goes away with N dollars and the other goes broke – this is the gambler's ruin. We ask:

What is the probability that Mike loses his fortune as a function of p and q?

Thinking optimistically, let W_I be the event that Mike wins the game if he starts with an initial fortune I

and define

$$P_I = \mathbb{P}(W_I).$$

To understand the function P_I, let's condition on the event F that Mike wins the first flip. In this case, the law of total probability gives

$$\begin{aligned} P_I &= \mathbb{P}(W_I|F)P(F) + \mathbb{P}(W_I|F^c)P(F^c) \\ &= \mathbb{P}(W_I|F)p + \mathbb{P}(W_I|F^c)q \\ &= p\mathbb{P}(W_I|F) + q\mathbb{P}(W_I|F^c) \end{aligned}$$

where we have used the fact that $\mathbb{P}(F) = p$ and $\mathbb{P}(F^c) = \mathbb{P}(\text{loses first flip}) = q$. To understand the conditional probabilities $\mathbb{P}(W_I|F)$, we see that at the start of the second flip, the only information that is retained is the amount of money that Mike has because the first trial is independent of all others. This is the nature of independent trials and is often given the name memorylessness (or called the Markov property). So, if Mike starts with an initial fortune of I dollars and it is known that he wins on the first flip, he is given a dollar so he then has $I+1$ dollars. Since whether or not he wins the game only depends on subsequent flips, we have

$$\mathbb{P}(W_I|F) = \mathbb{P}(W_{I+1}) = P_{I+1}$$

and similarly,

$$\mathbb{P}(W_I|F^c) = \mathbb{P}(W_{I-1}) = P_{I-1}.$$

All together, we obtain the equation

$$P_I = pP_{I+1} + qP_{I-1}$$

for $I = 1, 2, \ldots, N-1$. Upon recalling that $p + q = 1$, we can rewrite the above equation as

$$(p+q)P_I = pP_I + qP_I = pP_{I+1} + qP_{I-1}$$

or, equivalently,

$$P_{I+1} - P_I = \frac{q}{p}(P_I - P_{I-1}) \tag{4.2}$$

which holds for all $I = 1, 1, N-1$. Also, for $I = 0$, it is clear that Mike loses and so $P_0 = 0$ and, similarly, $P_N = 1$. Thus we obtain a so-called boundary value problem for the difference equation (4.2): Find P_I that satisfies

$$\begin{cases} P_{I+1} - P_I = \frac{q}{p}(P_I - P_{I-1}) \quad I = 1, 2, \ldots, N-1 \\ P_0 = 0 \\ P_N = 1 \end{cases} \tag{4.3}$$

To solve the above boundary value problem, we first observe that, for $I = 1$,

$$P_2 - P_1 = \frac{q}{p}(P_1 - 0) = \frac{q}{p}P_1$$

where we have used the boundary condition that $P_0 = 0$. Pushing this information forward, we see that

$$P_3 - P_2 = \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1$$

and, for any $I = 1, 2, N$, we have

$$P_I - P_{i-1} = \left(\frac{q}{p}\right)^{I-1} P_1.$$

Using the idea of telescoping sums, we observe that, for each $I = 1, 2, \ldots,$

$$\begin{aligned} P_I &= P_I - P_1 + P_1 - P_0 \\ &= P_I - P_{I-1} + P_{I-1} - P_{I-2} + \cdots - P_2 + P_2 - P_1 + P_1 - P_0 \\ &= \sum_{i=1}^{I} (P_i - P_{i-1}) \\ &= \sum_{i=1}^{I} \left(\frac{p}{q}\right)^{i-1} P_1 \\ &= P_1 \sum_{i=1}^{I} \left(\frac{p}{q}\right)^{i-1}. \end{aligned}$$

Recalling the formula for partial sums of the geometric series,

$$\sum_{i=1}^{I} r^{i-1} = \begin{cases} \frac{1-r^i}{1-r} & r \neq 1 \\ I & r = 1 \end{cases},$$

we have

$$P_I = P_1 \times \begin{cases} \frac{1-(q/p)^I}{1-q/p} & p \neq q \\ I & p = q = 1/2 \end{cases}$$

which is valid for $I = 0, 1, \ldots, N$. To determine P_I completely, it remains only to find the value of P_1 and we do this by using the boundary condition $P_N = 1$. Thus

$$1 = P_N = P_1 \begin{cases} \frac{1-(q/p)^N}{1-q/p} & p \neq q \\ N & p = q = 1/2 \end{cases}$$

or

$$P_1 = \begin{cases} \frac{1-q/p}{1-(q/p)^N} & p \neq q \\ \frac{1}{N} & p = q = 1/2. \end{cases}$$

Thus, the solution to our boundary value problem is

$$P_I = \begin{cases} \frac{1-(q/p)^I}{1-(q/p)^N} & p \neq q \\ \frac{I}{N} & p = q = 1/2. \end{cases}$$

Since Gambler's ruin (or Mike going broke) is the complement of the event W_I, we have

$$\mathbb{P}(\text{Mike, with an initial fortune } I, \text{ goes broke}) = 1 - P_I = \begin{cases} \frac{(q/p)^I - (q/p)^N}{1-(q/p)^N} & p \neq q \\ \frac{N-I}{N} & p = q = 1/2 \end{cases}$$

In Figure 4.1, I have illustrated this probability $1 - P_I$ for various ratios of q/p with $N = 10$. In the case that $q/p = 3$ (i.e., $p = 1/4$ and $q = 3/4$), the figure shows that Mike has a high probability of going broke unless he walks in with $I = 9$ or 10 dollars. In the case that $q/p = 1/3$ (i.e., $p = 3/4$ and $q = 1/4$), the figure shows that Mike has a low chance of going broke for $I \geq 2$.

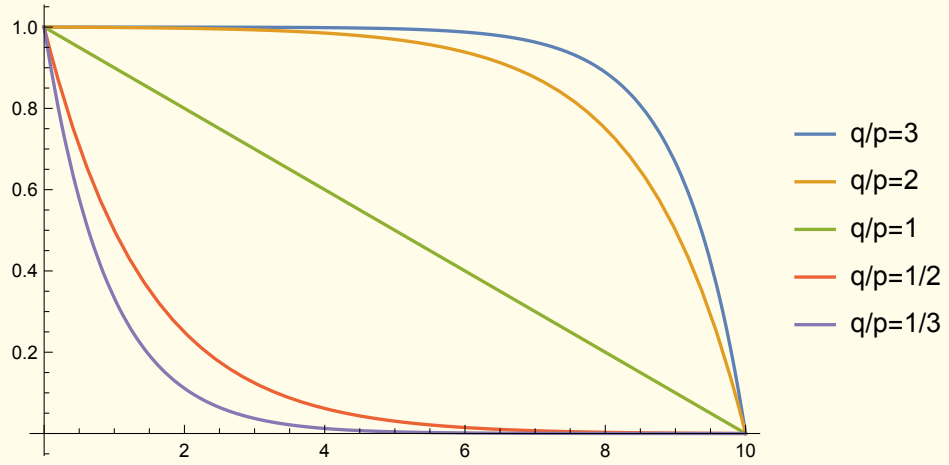


Figure 4.1: Graphs of $1 - P_I$ for various ratios of q/p.

Finally, I want to point something out to those who have interest in probability's connection to partial differential equations and physics. Boundary value problems of the form (4.3) are often viewed as boundary value problems for the discrete Laplace's equation,

$$\Delta u = 0$$

where Δ is called the discrete Laplacian on $\mathbb{N}$ and, in this setting, is defined by

$$\Delta u(i) = u(i) - pu(i+1) - qu(i-1)$$

for integers i. Here, (4.3) is

$$\begin{cases} \Delta P = 0 & \text{on } \mathbb{N}_+ = \{1, 2, \ldots, N-1\} \\ P = 0 & \text{at } (\partial \mathbb{N}_+)_L = \{0\}, \text{ the left boundary of } \mathbb{N}_+ \\ P = 1 & \text{at } (\partial \mathbb{N}_+)_R = \{N\}, \text{ the right boundary of } \mathbb{N}_+. \end{cases}$$

These types of boundary value problems for Laplace's equation have many important applications in physics. For example, let D be a region in $\mathbb{R}^3$ (think of a room) and ∂D be its boundary (think of the walls of the room), if one specifies a temperature profile $f(x, y, z)$ on the boundary of the room, i.e., $f : \partial D \to [0, \infty)$, the steady-state temperature in the interior of the room $u(x, y, z)$ is a solution to the boundary value problem

$$\begin{cases} \Delta u = 0 & \text{in } D \\ u = f & \text{on } \partial D. \end{cases}$$

Here, Δ is the (continuous) Laplacian on $\mathbb{R}^3$ defined by

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

In other words, if you know the temperature at the walls of a room (which is, say, determined by what's going on outside), you can determine the temperature at every point in the room by solving the above boundary value problem. It turns out that the above problem can be solved using probability theory (very much akin to Gambler's ruin) and this connection underlies the connection between heat transfer/temperature and Brownian motion/random walk – which we will talk about later in the semester. If you'd like to read more about this, please take a look at the text Green, Brown, and Probability by Kai Lai Chung.

Exercise 4.10: Gambler's Ruin – playing against a casino

Suppose that you are gambling against an infinitely rich adversary and, at each stage, you either win or lose \$1 with probability p and $q = 1 - p$, respectively. We shall assume that the games/stages are independent and so represent independent trials. Show that the probability you eventually go broke is

$$P(\text{Going Broke}) = \begin{cases} 1 & p \leq 1/2 \\ \left(\frac{q}{p}\right)^I & p > 1/2 \end{cases}$$

where I is your initial fortune. There are a couple of different ways to solve this, but you should to back through the steps of the above example (working with P_I and replace the boundary condition $P_N = 1$ with

$$\lim_{I \to \infty} P_I = 1.$$

Exercise 4.11: Alice's Random Walk

Suppose that a "random walker" named Alice walks along the integers,

$$\mathbb{Z} = \{\ldots, -4, -3, -2, -1, 0, 1, 2, 3, 4, \ldots\}.$$

Starting out at position 0, suppose that Alice flips a fair coin and then walks to the left one unit if the coin is "tails" and to the right one unit if the coin is "heads". No matter where Alice is after the first step, she then flips the coin (independently) and moves one step to the left if the coin is "tails" and one step to the right if it is "heads". She repeats this process n times. Our goal in this problem is to understand where Alice is after n steps and then compute the probability of her getting/being there.
To this end, denote by S_n Alice's position after n steps and write

$$p_n(k) = \mathbb{P}(S_n = k)$$

for each position$k \in \mathbb{Z}$ and number of steps $n = 1, 2, \ldots$. We note that $p_n(k)$ tells us the probability that Alice is standing at position k after n steps. Observe that the event $\{S_1 = 1\}$ is precisely the event that the coin came up heads after the first step and so

$$p_1(1) = \mathbb{P}(S_1 = 1) = \mathbb{P}(H) = \frac{1}{2}.$$

Similarly,

$$p_1(-1) - \mathbb{P}(S_1 = -1) = \mathbb{P}(T) = \frac{1}{2}.$$

Since she has no chance of being at any other position after 1 step, we have

$$p_1(k) = \begin{cases} \frac{1}{2} & k = \pm 1 \\ 0 & \text{else.} \end{cases}$$

Determining $p_2(k)$ is slightly more involved. First, we see that the event $\{S_2 = 2\}$ is precisely the event that the coin came up heads twice. Since these flips are independent, we have

$$p_2(2) = \mathbb{P}(S_2 = 2) = \mathbb{P}(HH) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

and similarly

$$p_2(-2) = \mathbb{P}(S_2 = -2) = \mathbb{P}(TT) = \frac{1}{4}.$$

Thinking about Alice's other possible sequence of movements (called *sample paths*), we see that, after two steps, the only other possible position is 0. This happens if she first went to 1 and then back to 0 (heads then tails) or first went to -1 and then back to 0 (tails then heads). We note that each of these two distinct sample paths has probability $1/2^2$ and so

$$p_2(0) = \mathbb{P}(S_2 = 0) = \mathbb{P}(\{HT, TH\}) = \mathbb{P}(HT) + \mathbb{P}(TH) = \frac{1}{2^2} + \frac{1}{2^2} = \frac{1}{2}.$$

Thus

$$p_2(k) = \begin{cases} \frac{1}{4} & k = \pm 2 \\ \frac{1}{2} & k = 0 \\ 0 & \text{else.} \end{cases}$$

1. Please determine $p_3(k)$ for all $k \in \mathbb{Z}$. As you suspect, $p_3(k)$ should only be non-zero for a small number of k's.

2. For each n, argue that whenever $k \in \mathbb{Z}$ is a position for which $p_n(k) > 0$ (i.e., that it's actually possible for Alice to be there after n steps),

$$p_n(k) = N(n, k)\left(\frac{1}{2^n}\right)$$

   where $N(n, k)$ denotes the number of distinct ways that Alice can walk from 0 to k in n steps.

3. By viewing this as a counting problem, compute $N(n, k)$.

4. What is the probability that after $n = 2m$ steps, Alice has returned to the origin?

Chapter 5

Random Variables

As we have seen, when modeling experiments (and their outcomes and events), it is often helpful to encode certain information about the outcomes using numerical values, i.e., we can take measurements. For example, we can understand the outcomes of a single coin flip ("H" or "T") by assigning the number $X = 1$ to H and the number $X = 0$ to T. In the more complicated setting of n coin flips (cf. Example **??**), we can study the experiment's outcomes by assigning numerical values to various observations. For example, we could encode information of just the last flip by assigning $X = 1$ if the last flip is H and $X = 0$ otherwise. We could instead encode information based on all n flips, for example, by counting the number of H observe in n flips. In the Dart Throwing Example, we could measure the distance between the landing position and the bullseye. These are all examples of the following.

Definition 5.1. *Let Ω be a sample space. Any*[1] *real-valued function on Ω, i.e., $X : \Omega \to \mathbb{R}$, is said to be a random variable. The range of X is the set*

$$R(X) = \{X(\omega) \in \mathbb{R} : \omega \in \Omega\}.$$

That is, for a random variable X, $R(X)$ is the set of all real numbers x which can be expressed as $x = X(\omega)$ for some $\omega \in \Omega$.

Example 5.1: A Very Simple Example

Consider the experiment of flipping a single coin, $\Omega = \{H, T\}$ and, on Ω, define the random variable

$$X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$$

As we discussed, this random variable encodes "heads or not" about the outcomes of this simple experiment. Here $R(X) = \{0, 1\}$.

Example 5.2: Flip and Roll

Consider an experiment in which we flip one coin and roll one die. We can take

$$\Omega = \{(c, d) : c = H \text{ or } T, d = 1, 2, \ldots, 6\}$$

[1] If you continue with probability beyond this course, you will learn that this definition actually needs to be slightly more restrictive and random variables are required to be so-called measurable functions. The concept measurablility is a delicate one and won't ever be an issue for us.

On this Ω, we can define a random variable X by putting

$$X(\omega) = X(c,d) = \begin{cases} 10 & c = H \text{ and } d \text{ is even} \\ -2 & c = H \text{ and } d \text{ is odd} \\ 1/2 & c = T \text{ and } d = 1,2 \\ 0 & c = T \text{ and } d = 3,4,5,6 \end{cases}$$

for $\omega = (c,d) \in \Omega$. This random variable could represent, for instance, the winnings associated with a player flipping the coin and rolling the die. Here $R(X) = \{-2, 0, 1/2, 10\}$.

Example 5.3: Flips to First Heads

Consider an experiment in which we flip one coin over and over again until the first heads appears. this should reference an earlier example. As before, we can describe the outcomes of this experiment by

$$\Omega = \{H, TH, TTH, TTTH, TTTTH, \dots\}$$

For this experiment, we can count the number of times the coin is flipped until heads appears (and the game is stopped). This is the random variable $N : \Omega \to \mathbb{R}$ defined by

$$N(\omega) = \begin{cases} 1 & \omega = H \\ 2 & \omega = TH \\ 3 & \omega = TTH \\ \vdots & \vdots \end{cases}$$

Actually, we can give this definition a little bit more elegantly by writing

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$$

where, for $k = 1, 2, \dots,$

$$\omega_k = \underbrace{TTT\cdots T}_{k-1} H.$$

In this presentation,

$$N(\omega_k) = k$$

for $k = 1, 2, \dots.$ Here,

$$R(N) = \{1, 2, 3, \dots\} = \mathbb{N}_+.$$

Example 5.4: Distance on a Dart Board

Consider throwing a dart at a dartboard of radius 1 unit. As we discussed before, we can represent the sample space for this experiment by

$$\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

On Ω, lets construct a random variable that measures the distance from the landing position of the dart to the bullseye, $(0,0)$. This is $D : \Omega \to \mathbb{R}$ defined by

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

for $\omega = (x,y) \in \Omega$. It is clear that the range of D is all real numbers from 0 to 1, i.e., $R(D) = [0,1]$.

Remark 5.2. Notice that the random variables of Examples 5 and 5 have ranges that are subsets of the integers $\mathbb{Z}$. In the case that a random variable X has $R(X) \subseteq \mathbb{Z}$, we say that X is *integer valued*. It turns out that integer-valued random variables are often the easiest to study (and it has to do with the fact that Fourier analysis is so powerful when studying functions on $\mathbb{Z}$). Observe that the random variables in Examples 5 and 5 are not integer valued. Referencingissue.

Armed with the notion of random variables on a sample space, we can ask about events associated to those random variables. For example, we can ask about the event that a random variable X takes on the value 1; this is the event

$$\{\omega \in \Omega : X(\omega) = 1\}.$$

Similarly, we could ask about the event that the random variable is positive; this is the event

$$\{\omega \in \Omega : X(\omega) > 0\}.$$

We shall employ the common notation[2] that, for a set of real numbers I,

$$\{X \in I\} := \{\omega \in \Omega : X(\omega) \in I\}.$$

This notation takes some getting used to because, while I is a set of real numbers, $\{X \in I\}$ is an event, that is, a subset of Ω. In the case that I is an interval of the form $[a, b]$, we write

$$\{a \leq X \leq b\} = \{\omega \in \Omega : a \leq X(\omega) \leq b\}.$$

The events $\{a < X \leq b\}$, $\{a \leq X < b\}$, $\{a < X < b\}$, $\{X \leq b\}$, etc., are defined analogously. Of course, now that we have events, we can start discussing probabilities of these events. You should make note that everything we've done so far in this chapter has not made reference to probability. It is high time to do so. Let Ω be a sample space equipped with a probability measure $\mathbb{P}$ and let X be a random variable on Ω. Understanding the likelyhood of the values of X is precisely the question of understanding the probabilities

$$\mathbb{P}(X \in I) := \mathbb{P}(\{X \in I\})$$

for various sets of real numbers $I \subseteq \mathbb{R}$. As we will see over time, these probabilities for one particular collection of intervals is particularly illuminating.

Definition 5.3. *Let X be a random variable on a sample space Ω equipped with probability measure $\mathbb{P}$. The cumulative distribution function of X is the function $F_X : \mathbb{R} \to \mathbb{R}$ given by*

$$F_X(x) = \mathbb{P}(X \leq x)$$

defined for each $x \in \mathbb{R}$. We will refer to F_X as the CDF of X; this is sometimes also referred to as X's distribution – though this term is used for several other things.

Example 5.5: The Bernoulli Random Variable

Let's return to the example of flipping a single coin, $\Omega = \{H, T\}$, and let's assume that this is, in general, a biased coin with probability $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1 - p$ for some $0 \leq p \leq 1$. Let X be the random variable

$$X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}.$$

This is an example of a Bernoulli random variable named in honor of Jacob Bernoulli. We can now compute

[2] If you are familiar with the concept of preimage, the event $\{X \in I\} = X^{-1}(I)$ is precisely the preimage of I under X.

the probabilities of various events involving X. For example, since $\{X = 1\} = \{H\}$,

$$\mathbb{P}(X = 1) = \mathbb{P}(H) = p.$$

Similarly, $\mathbb{P}(X = 0) = \mathbb{P}(T) = q$ and we observe that, for any number x which is not 0 or 1, $\mathbb{P}(X = x) = \mathbb{P}(\varnothing) = 0$. As you can probably guess given the simplicity of this random variable, its cumulative distribution function is easy to compute. For $x < 0$, we have

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\varnothing) = 0.$$

For $0 \leq x < 1$,

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X = 0) = q$$

and, for $x \geq 1$,

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X = 0, 1) = 1.$$

This is illustrated in Figure 5.1.

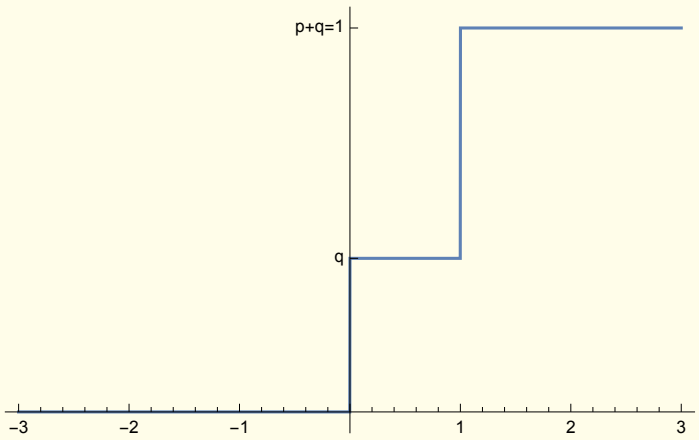


Figure 5.1: CDF for a Bernoulli random variable

Example 5.6: Probabilities of Flip and Roll

In the coin flip/die roll experiment of Example 5, let's assume additionally that the coin is fair, the die is perfect, and the flip and roll happen independently. With these assumptions, the probability measure of interest is the uniform measure

$$\mathbb{P}(A) = \frac{\#(A)}{\#(\Omega)} = \frac{\#(A)}{12}$$

on the 12-element sample space $\Omega = \{(c, d) : c = H \text{ or } T,\ d = 1, 2, \ldots, 6\}$. Let's consider the random variable X defined in Example 5 which could represent the winnings associated to a player flipping the coin and rolling the die. We could compute the probability that these winnings are positive. We could ask, what is the probability that the winnings are positive? Looking at the definition of X, we observe that

$$\{\text{Positive Winnings}\} = \{X > 0\} = \{\omega : X(\omega) = 1 \text{ or } 10\} = \{(H, 2), (H, 4), (H, 6), (T, 1), (T, 2)\}.$$

Therefore

$$\mathbb{P}(\text{Positive Winnings}) = \mathbb{P}(X > 0) = \frac{\#\{(H, 2), (H, 4), (H, 6), (T, 1), (T, 2)\}}{12} = \frac{5}{12}.$$

Example 5.7: Probability to First Heads

Let's return to our experiment where we flip a coin over and over until the first appearance of "heads". We saw that this sample space could be realized by

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$$

where, for $k = 1, 2, 3, \dots$,

$$\omega_k = \underbrace{TTT\cdots T}_{k-1} H.$$

If we assume that the coin is biased (with each flip coming up heads with probability $0 \leq p < 1$ and tails with probability $q = 1 - p$) and we assume that the flips are independent, we can compute the probability of the outcomes by

$$\mathbb{P}(\omega_k) = \mathbb{P}(\{\underbrace{TTT\cdots T}_{k-1} H\}) = q^{k-1}p$$

As before, consider the random variable N which counts the number of flips until heads appears, i.e., that for which $N(\omega_k) = k$ for $k = 1, 2, \dots$ and whose range is $R(N) = \mathbb{N}_+$. We can compute various probabilities associated to this random variable N. For example, for the event $\{N = k\}$ that it takes exactly k flips for heads to appear (for $k \in \mathbb{N}_+$), we have

$$\mathbb{P}(N = k) = \mathbb{P}(\omega_k) = q^{k-1}p$$

For the event $\{N \leq 3\}$ that it takes at most three flips to see heads, we have

$$\mathbb{P}(N \leq 3) = \mathbb{P}(\{\omega_1, \omega_2, \omega_3\}) = \mathbb{P}(\omega_1) + \mathbb{P}(\omega_2) + \mathbb{P}(\omega_3) = p + qp + q^2p = p(1 + q + q^2).$$

Continuing in this way, we see that

$$\begin{aligned}
\mathbb{P}(N \leq n) &= p + qp + q^2p + \cdots + q^{n-1}p \\
&= p(1 + q + q^2 + \cdots + q^{n-1}) \\
&= p\left(\sum_{k=1}^{n} q^{k-1}\right) \\
&= p\frac{1 - q^n}{1 - q} \\
&= 1 - q^n
\end{aligned}$$

whenever n is a natural number. In the above I have used the assumption that $p < 1$ so that $1 - q = p < 1$ and the formula

$$\begin{aligned}
(1 - x)\sum_{k=1}^{n} x^{k-1} &= (1 - x)(1 + x + x^2 + \cdots + x^{n-1}) \\
&= 1 - x + x - x^2 + x^2 - \cdots - x^{n-1} + x^{n-1} - x^n = 1 - x^n
\end{aligned}$$

useful in the computation of geometric series and its partial sums. If we recall the definition of the so-called floor function,

$$\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \leq x\},$$

we can express the cumulative distribution function of N by

$$F_N(x) = \begin{cases} 0 & x < 0 \\ 1 - q^{\lfloor x \rfloor} & x \geq 0 \end{cases}$$

for $x \in \mathbb{R}$ where we have used the fact that N can only assume integer values and so

$$\mathbb{P}(N \leq x) = \mathbb{P}(N \leq n = \lfloor x \rfloor)$$

for $x \geq 0$. This cumulative distribution function is illustrated in Figure 5.2.

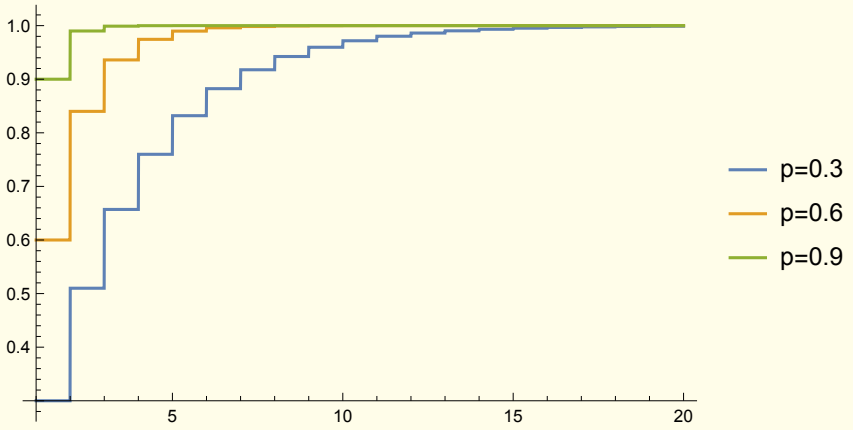


Figure 5.2: CDF of Geometric random variables various values of p.

Exercise 5.1: Properties of the CDF

Looking back through the cumulative distribution functions of this section, we observe the following properties: They appear to be non-decreasing. They also appear to have limits of 0 at $-\infty$ and 1 at infinity. In this exercise, you will show that this is, in fact, always the case. Let X be a random variable on a sample space X equipped with probability measure $\mathbb{P}$.

1. By considering the events $\{X \leq x_1\}$, $\{X \leq x_2\}$ for $x_1 \leq x_2$, use the monotonicity of probability to show that
$$F_X(x_1) \leq F_X(x_2)$$
whenever $x_1 \leq x_2$. This is precisely what it means to say that F_X is non-decreasing.

2. Use the continuity of probability to show that
$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \text{and} \qquad \lim_{x \to \infty} F_X(x) = 1.$$
Hint: For the second limit, notice that the collection $\{X \leq n\}_{n=1,2,\ldots}$ is a nested increasing collection of events whose union is Ω.

5.1 Expectation (on countable sample spaces)

Now that we are familiar with random variables and computing probabilities associated to them, we are ready to introduce the expectation. As we will see, this object/quantity will be of prime interest and utility for us throughout the remainder of the course. A complete and correct introduction of the expectation requires knowledge of a subject called measure theory (a subdiscipline of analysis) and so we will not do that here. In the special case that our underlying sample space Ω is countable (either finite or countably infinite), we can define the expectation precisely as it only involves summation (either finite sums or infinite series) as we will see. Fortunately, the majority of examples we have seen so far are posed on countable sample spaces and so the theory treated here will be of broad

applicability to us.

To this end, consider a countable sample space Ω of the form

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}.$$

which is equipped with a probability measure $\mathbb{P}$. We recall (Proposition 2.7) that $\mathbb{P}$ is characterized by its values on each outcome, i.e., the values $p_n = \mathbb{P}(\omega_n)$ for $n = 1, 2, \dots$ which necessarily satisfy $\sum_n p_n = 1$. On Ω, we consider a random variable $X : \Omega \to \mathbb{R}$ which has range

$$R(X) = \{X(\omega) : \omega \in \Omega\} = \{X(\omega_1), X(\omega_2), X(\omega_3), \dots\} \tag{5.1}$$

which is necessarily countable. With this landscape as our background, we introduce the expectation:

Definition 5.4. *Let Ω be a countable sample space equipped with a probability measure $\mathbb{P}$ and let X be a random variable on Ω.*

- *In the case that the series*
$$\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)$$
converges absolutely, we define the expectation of X to be the sum of the series and we denote it by $\mathbb{E}(X)$. In other words,
$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)$$
provided that
$$\sum_{\omega\in\Omega} |X(\omega)|\,\mathbb{P}(\omega) < \infty.$$
The expectation of X is also called the expected value of X and the mean of X.

- *More generally, if φ is real-valued function of a real variable (i.e., $\varphi : \mathbb{R} \to \mathbb{R}$), the expectation of the random variable $\varphi(X)$ is defined by*
$$\mathbb{E}(\varphi(X)) = \sum_{\omega\in\Omega} \varphi(X(\omega))\mathbb{P}(\omega)$$
provided that the defining series above converges absolutely.

Remark 5.5. The condition that the series $\sum X(\omega)\mathbb{P}(\omega)$ is absolutely convergent will be automatic for most random variables you will encounter in this course. For example, if Ω is finite, then absolute convergence is immediate because all sums involved are finite. Unless you are explicitly asked to verify absolute convergence (in this course), you can take it for granted.

Remark 5.6. The absolute summation $\sum |X(\omega)|\,\mathbb{P}(\omega)$ is precisely the series defining $\mathbb{E}(|X|)$, i.e, $\mathbb{E}(\varphi(X))$ when $\varphi(x) = |x|$. Since this is a series whose terms/summands are non-negative, this series either converges or diverges to infinity (by the monotone convergence theorem). In either case, we can make sense of the expectation $\mathbb{E}(|X|)$ as either a finite number or infinity. With this interpretation, we can say that $\mathbb{E}(X)$ is defined whenever $\mathbb{E}(|X|) < \infty$. More generally, $\mathbb{E}(\varphi(X))$ is defined whenever $\mathbb{E}(|\varphi(X)|) < \infty$.

Remark 5.7. When Ω is finite, the manner in which all summations are computed is unambiguous. They are finite sums! When Ω is infinite (but still countable), one can perform the summation of the series via the enumeration $\Omega = \{\omega_1, \omega_2, \dots, \}$ by computing

$$\sum_{n=1}^{\infty} X(\omega_n)\mathbb{P}(\omega_n) = \lim_{N\to\infty} \sum_{n=1}^{N} X(\omega_n)\mathbb{P}(\omega_n).$$

The astute reader might ask: If I used a different enumeration of Ω, would this series still converge (and to the same number)? If you recall the intricacies of series and, in particular, the fascinating behavior of the rearrangements of conditionally convergent series, this is a very natural question (c.f., [**?**, Theorem 3.54]). Fortunately, there is a result that says that, when a series is absolutely convergent, this sum is unique and the order in which the sum is taken is immaterial (c.f., [**?**, Theorem 3.55]). This is precisely why we ask for absolute convergence.

Example 5.8: A Simple Example for Expectation

Consider the single coin flip experiment $\Omega = \{H, T\}$ with probability measure $\mathbb{P}$ with $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1 - p$. Define

$$X(\omega) = \begin{cases} 1 & \omega = H \\ -1 & \omega = T \end{cases}$$

In this case,

$$\begin{aligned} \mathbb{E}(X) &= \sum_{\omega \in \{H,T\}} X(\omega)\mathbb{P}(\omega) = X(H)\mathbb{P}(H) + X(T)\mathbb{P}(T) = 1 \cdot p + (-1) \cdot q \\ &= p - q \end{aligned}$$

Example 5.9: Expectation for Flip and Roll

Consider the coin flip/die roll of Examples 5.2 and 5.6. Given that all 12 outcomes are equally likely, i.e, $\mathbb{P}(\omega) = 1/12$ for each outcomes ω, we have

$$\begin{aligned} \mathbb{E}(X) &= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) \\ &= \sum_{\omega \in \Omega} X(\omega)\frac{1}{12} \\ &= \frac{1}{12}\Big(X(H,1) + X(H,2) + X(H,3) + X(H,4) + X(H,5) + X(H,6) \\ &\qquad + X(T,1) + X(T,2) + X(T,3) + X(T,4) + X(T,5) + X(T,6)\Big) \\ &= \frac{1}{12}\Big((-2) + 10 + (-2) + 10 + (-2) + 10 \\ &\qquad + \frac{1}{2} + \frac{1}{2} + 0 + 0 + 0 + 0\Big) \\ &= \frac{25}{12} \end{aligned}$$

The following proposition captures two basic facts of expectation. The first says that the expectation is positivity preserving (in that it takes non-negative random variables to non-negative numbers). The second is the property that the expectation is linear.

Proposition 5.8. *Let X be a random variable on a sample space Ω with probability measure $\mathbb{P}$.*

1. *If X is non-negative, i.e., $X(\omega) \geq 0$ for all $\omega \in \Omega$, then $\mathbb{E}(X) \geq 0$.*
2. *If X and Y are random variables with $\mathbb{E}(|X|) < \infty$ and $\mathbb{E}(|Y|) < \infty$, then, for any constants α and β,*

$$\mathbb{E}(|\alpha X + \beta Y|) = \sum_{\omega \in \Omega} |\alpha X(\omega) + \beta Y(\omega)|\, \mathbb{P}(\omega) < \infty$$

and

$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y).$$

Remark 5.9. When X is non-negative, $X = |X|$ and so, in view of Remark 5.6, $\mathbb{E}(X) = \mathbb{E}(|X|)$ has a meaning regardless of whether or not the defining series converges. For this reason, we don't require $E(|X|) < \infty$ for the hypothesis in Item 1 for, in the case that $\mathbb{E}(X) = \mathbb{E}(|X|) = \infty$, we simply obtain the (very reasonable) inequality $0 \leq \infty$.

Proof. The proposition is stated and is, in fact, true regardless of whether or not Ω is countable (provided that $\mathbb{E}$ is defined correctly). Still, as we have only the definition for $\mathbb{E}$ in the case that Ω is countable, we shall prove things only in that case.

To see Item 1, let's first write

$$S_N = \sum_{n=1}^{N} X(\omega_n)\mathbb{P}(\omega_n)$$

for $N \in \mathbb{N}_+$. In view of Remark 5.7, these are the partial sums of the series defining $\mathbb{E}(X)$. Since, for each n, $X(\omega_n) \geq 0$ and $\mathbb{P}(\omega_n) \geq 0$, it is clear that $S_N \geq 0$ for each $N \in \mathbb{N}_+$. Consequently,

$$\mathbb{E}(X) = \sum_{\omega} X(\omega)\mathbb{P}(\omega) = \lim_{N\to\infty} \sum_{n=1}^{N} X(\omega_n)\mathbb{P}(\omega_n) = \lim_{N\to\infty} S_N \geq 0.$$

For Item 2, we shall prove only the special case in which $\beta = 0$; the full result is a homework exercise. To this end, we first observe that, for any $\omega \in \Omega$, $|\alpha X(\omega)| = |\alpha|\,|X(\omega)|$ and therefore

$$\begin{aligned}
\sum_{\omega\in\Omega} |\alpha X(\omega)|\,\mathbb{P}(\omega) &= \sum_{\omega} |\alpha|\,|X(\omega)|\,\mathbb{P}(\omega) \\
&= \lim_{N\to\infty} \sum_{n=1}^{N} |\alpha|\,|X(\omega_n)|\,\mathbb{P}(\omega_n) \\
&= \lim_{N\to\infty} |\alpha| \sum_{n=1}^{N} |X(\omega_n)|\,\mathbb{P}(\omega_n) \\
&= |\alpha| \left(\lim_{N\to\infty} \sum_{n=1}^{N} |X(\omega_n)|\,\mathbb{P}(\omega_n) \right) \\
&= |\alpha| \left(\sum_{\omega\in\Omega} |X(\omega)|\,\mathbb{P}(\omega) \right) \\
&= |\alpha|\,\mathbb{E}(|X|) < \infty
\end{aligned}$$

where we have invoked our hypothesis that $\mathbb{E}(|X|) < \infty$. Since absolute convergence guarantees convergence, we have

$$\begin{aligned}
\mathbb{E}(\alpha X) &= \sum_{\omega\in\Omega} \alpha X(\omega)\mathbb{P}(\omega) = \lim_{N\to\infty} \sum_{n=1}^{N} \alpha X(\omega_n)\mathbb{P}(\omega_n) \\
&= \lim_{N\to\infty} \alpha \sum_{n=1}^{N} X(\omega_n)\mathbb{P}(\omega_n) = \alpha \left(\lim_{N\to\infty} \sum_{n=1}^{N} X(\omega_n)\mathbb{P}(\omega_n) \right) = \alpha \left(\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) \right) = \alpha\mathbb{E}(X),
\end{aligned}$$

as desired. □

Exercise 5.2: Finishing the proof of Proposition 5.8

1. Let X and Y be random variables on a common countable sample space Ω with probability measure $\mathbb{P}$. Assume that

$$\mathbb{E}(|X|) = \sum_{\omega\in\Omega} |X(\omega)|\mathbb{P}(\omega) < \infty \qquad \text{and} \qquad \mathbb{E}(|Y|) = \sum_{\omega\in\Omega} |Y(\omega)|\mathbb{P}(\omega) < \infty,$$

Show that

$$\sum_{\omega\in\Omega} |X(\omega)+Y(\omega)|\mathbb{P}(\omega) < \infty$$

and

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + E(Y).$$

2. Use this result to complete the proof of Proposition 5.8. Hint: Think about applying the above result when X and Y are replaced by αX and βY, respectively.

3. Extend the result to a finite collection of random variables $X_1, X_2, \ldots, X_n$ all defined on the countable sample space Ω.

Exercise 5.3:

A biased coin is flipped three times independently, on each flip it comes up heads with probability p and tails with probability $q = 1-p$. Consider the random variable X which counts the number of heads that appear in these three consecutive flips (note that $R(X) = 0, 1, 2, 3$).

1. Write down a realization of the sample space Ω for this experiment.

2. Describe the probability measure $\mathbb{P}$ that incorporates the assumption that the three flips are independent where each flip is heads with probability p and tails with probability $q = 1-p$.

3. Write down a formula for X as a piecewise function on Ω.

4. Compute the expectation $\mathbb{E}(X)$.

5. Now, consider the random variable $Y = 2X - 3$ and write down a formula for Y as a piecewise function on Ω.

6. Compute the $\mathbb{E}(Y)$ directly (using the definition of expectation).

7. Compute $\mathbb{E}(Y)$ using linearity (Proposition 5.8) and your previous value for $\mathbb{E}(X)$; you can use the fact that $\mathbb{E}(1) = 1$ – a fact you will show in a later exercise.

At this point, it is natural to ask: What does the expected value of a random variable tell us? What does it measure? From its definition, you should think about the expectation as a weighted average of the values of the random variable which are weighted according to their probabilities. If we think of probabilities in the frequentest interpretation, we would then interpret this weighted average as a statistical average of the values of the random variable over many repeated independent trials of the same experiment. As we discussed, however, we have no real way to justify this frequentest interpretation a priori. At the end of the course, we study a big and celebrated result called the law of large numbers which will tell us that this abstractly defined quantity is precisely that gotten in the limit of such statistical averages.

We now introduce another a few more quantities, defined using the expectation, which are useful when studying random variables. These are the moments and variance and, as the expected value of mean can be interpreted as a weighted average, the variance is a measure of the deviation from this mean.

Definition 5.10 (The Variance and Moments)**.** *Let X be a random variable with mean $\mu = \mathbb{E}(X)$.*

1. *The variance of X is the quantity*

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \sum_{\omega\in\Omega} (X(\omega)-\mu)^2\mathbb{P}(\omega).$$

As the defining terms/summands for $\mathrm{Var}(X)$ are non-negative, this series either converges or diverges to

infinity (by the monotone convergence theorem). We say that X has finite variance if it converges and infinite variance when it does not.

2. *For a natural number n, the nth moment of X is quantity given by*

$$\mathbb{E}(X^n) = \sum_{\omega\in\Omega}(X(\omega))^n\mathbb{P}(\omega)$$

provided that $\mathbb{E}(|X|^n) < \infty$.

Example 5.10: Simple Variance I

Let's return to the example of a single flip of a biased coin: $\Omega = \{H, T\}$ with $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1-p$. For the random variable

$$X(\omega) = \begin{cases} 1 & \omega = H \\ -1 & \omega = T \end{cases},$$

we saw that $\mu = \mathbb{E}(X) = p - q$. With this, we compute

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}((X-\mu)^2) = (X(H) - (p-q))^2\mathbb{P}(H) + (X(T) - (p-q))^2\mathbb{P}(T) \\
&= (1-(p-q))^2 \cdot p + (-1-(p-q))^2 \cdot q \\
&= p(1 + (p-q)^2 - 2(p-q)) + q(1 + (p-q)^2 + 2(p-q)) \\
&= p(1+(p-q)^2) + q(1+(p-q)^2) - 2p(p-q) + 2q(p-q) \\
&= (p+q)(1+(p-q)^2) - 2(p-q)(p-q) \\
&= 1 + (p-q)^2 - 2(p-q)^2 \\
&= 1 - (p-q)^2
\end{aligned}$$

We now compute the moments of X. For a natural number n, we have

$$\mathbb{E}(X^n) = X(H)^n\mathbb{P}(H) + X(T)^n\mathbb{P}(T) = 1^n p + (-1)^n q = p + (-1)^n q. \tag{5.2}$$

Since $p + q = 1$, we see that

$$\mathbb{E}(X^n) = \begin{cases} 1 & n \text{ even} \\ p - q & n \text{ odd} \end{cases}.$$

Another example needed

The following amasses some basic facts about variance.

Proposition 5.11. *On a sample space Ω equipped with probability measure $\mathbb{P}$, let X be a random variable with mean $\mu = \mathbb{E}(X)$. Then*

$$\text{Var}(X) \geq 0.$$

Further, $\text{Var}(X)$ is finite if and only if $\mathbb{E}(X^2)$ is finite and, in this case,

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

Proof. Given that $\varphi(X) = (X-\mu)^2 \geq 0$ is non-negative, Proposition 5.8 ensures that $\text{Var}(X) = \mathbb{E}((X-\mu)^2) \geq 0$. To establish the second property, let us first observe that

$$\mathbb{E}(\mu^2) = \sum_{\omega\in\Omega}\mu^2\mathbb{P}(\omega) = \mu^2\sum_{\omega\in\Omega}\mathbb{P}(\omega) = \mu^2\mathbb{P}(\Omega) = \mu^2$$

where we have used the fact that Ω is countable to see that $\mathbb{P}(\Omega) = \sum_{\omega\in\Omega}\mathbb{P}(\omega) = 1$ by virtue of the additivity of $\mathbb{P}$. Also, observe that

$$\mathbb{E}(-2\mu X) = -2\mu\mathbb{E}(X) = -2\mu\cdot\mu = -2\mu^2$$

by virtue of Proposition 5.8. Now,

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2$$

from which we see that the expectation of the left side is finite if and only if the expectation of the right side is finite. Using the fact that $\mathbb{E}(\mu^2) = \mu^2$ and $\mathbb{E}(|X|) < \infty$ (since μ is defined), we may conclude that $\text{Var}(X) = \mathbb{E}((X-\mu)^2)$ is finite if and only if $\mathbb{E}(X^2)$ is finite. In this case, Proposition 5.8 and our previous computations show that

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

as was claimed. □

Example 5.11: Simple Variance II

Let's revisit the biased coin example where $\Omega = \{H, T\}$ with $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1-p$ and the random variable

$$X(\omega) = \begin{cases} 1 & \omega = H \\ -1 & \omega = T \end{cases}$$

Here, since $\mu = \mathbb{E}(X) = p - q$ and $\mathbb{E}(X^2) = 1$, we see that

$$\text{Var}(X) = \mathbb{E}(X^2) - \mu^2 = 1 - (p-q)^2$$

which is exactly what we found previously.

Exercise 5.4: Expected number of flips until first heads

Consider the coin flipping experiment presented in Example 5.7. There, a biased coin is flipped over and over independently until heads appears for the first time.. We considered the random variable N which counts the number of flips until heads appears and is defined by $N(\omega_k) = k$ where, for each $k \geq 1$,

$$\omega_k = \underbrace{TTT\cdots T}_{k-1} H \in \Omega$$

which had probability $\mathbb{P}(\omega_k) = q^{k-1}p$.

1. Compute the expectation $\mathbb{E}(N)$.
2. Compute the second moment $\mathbb{E}(N^2)$.
3. Use your previous results to compute the variance $\text{Var}(N)$.

Our next proposition shows that the expectation can be used to recover the probability measure $\mathbb{P}$.

Proposition 5.12. *Let Ω be a sample space with probability measure $\mathbb{P}$ and expectation $\mathbb{E}$. Then, for any event $A \subseteq \Omega$,*

$$\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A).$$

Remark 5.13. If you take a graduate course in probability (or a course in measure theory), you will see that the above the equation $\mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A)$ is that which is used to construct the expectation in the first place. That is, one uses a probability measure to construct the expectation and, with and expectation, you can always recover the probability measure.

Exercise 5.5: Non-Random Random Variables

Assume that Ω is a countable sample space and $\mathbb{P}$ is a probability measure on Ω.

1. In this setting, prove Proposition 5.12.
2. If you recall from the proof of Proposition 5.11, we established that the constant random variable $\omega \mapsto \mu^2$ had $\mathbb{E}(\mu^2) = \mu^2$. Show that, for any constant C,

$$\mathbb{E}(C) = C \qquad \text{and} \qquad \text{Var}(C) = 0.$$

In other words, a non-random random variable C has itself as its mean and zero variance. Hint: To show this, write $C = C \cdot \mathbb{1}_\Omega$ and apply Proposition 5.12.

Our next theorem is of huge practical, theoretical, and philosophical importance. While the full scope of these virtues won't immediately be forthcoming, you will immediately be able to appreciate how much it simplifies the task of computing expectation.

Theorem 5.14. *Let Ω be a countable sample space with probability measure $\mathbb{P}$ and let X be a random variable with range $R(X)$ which is necessarily countable in view of (5.1). Then the series $\sum_{k\in R(X)} k \cdot \mathbb{P}(X = k)$ is absolutely convergent if and only if the series $\sum_\omega X(\omega)\mathbb{P}(\omega)$ is absolutely convergent and, in this case,*

$$\mathbb{E}(X) = \sum_{k\in R(X)} k \cdot \mathbb{P}(X = k) \tag{5.3}$$

More generally, given $\varphi : \mathbb{R} \to \mathbb{R}$, the series $\sum_{k\in R(X)} \varphi(k)\mathbb{P}(X = k)$ is absolutely convergent if and only if the series $\sum_\omega \varphi(X(\omega))\mathbb{P}(\omega)$ and, in this case,

$$\mathbb{E}(\varphi(X)) = \sum_{k\in R(X)} \varphi(k)\mathbb{P}(X = k). \tag{5.4}$$

Proof. We remark that that statement concerning (5.3) is a special case of that concerning (5.4) (when $\varphi(x) = x$) and so we shall discuss the validity of the second statement. The assertion that the two series converge absolutely (or do not) together is somewhat beyond the scope of what I want to communicate to you in this text (though I encourage you to try to prove it). Thus, I will make the assumption that the series

$$\sum_{\omega\in\Omega} \varphi(X(\omega))\mathbb{P}(\omega) \qquad \text{and} \qquad \sum_{k\in R(X)} \varphi(k) \cdot \mathbb{P}(X = k)$$

both converge absolutely. With this assumption, my job is to show that these series are, in fact, equal and so

$$\mathbb{E}(\varphi(X)) = \sum_{\omega\in\Omega} \varphi(X(\omega))\mathbb{P}(\omega) = \sum_{k\in R(X)} \varphi(k) \cdot \mathbb{P}(X = k)$$

For each $k \in R(X)$, let's define

$$A_k = \{X = k\} = \{\omega \in \Omega : X(\omega) = k\}.$$

and observe that

$$\Omega = \bigcup_{k\in R(X)} A_k$$

Correspondingly, a sum over outcomes $\omega \in \Omega$ can be iterated by summing first over those $\omega \in A_k$ (for each k) and then summing the result over those $k \in R(X)$. In other words,

$$\sum_{\omega\in\Omega} \varphi(X(\omega))\mathbb{P}(\omega) = \sum_{k\in R(X)} \sum_{\omega\in A_k} \varphi(X(\omega))\mathbb{P}(\omega).$$

This is technically a rearrangement argument and so one has to be careful but everything works out because of absolute convergence. To compute the inner sum, first observe that, for each $\omega \in A_k$, $X(\omega) = k$ and therefore

$$\sum_{\omega \in A_k} \varphi(X(\omega))\mathbb{P}(\omega) = \sum_{\omega \in A_k} \varphi(k) \cdot \mathbb{P}(\omega) = \varphi(k) \sum_{\omega \in A_k} P(\omega) = \varphi(k) \cdot \mathbb{P}(A_k)$$

where we have used countable additivity and the fact that A_k is the union of its outcomes. Consequently,

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} \varphi(X(\omega))\mathbb{P}(\omega) = \sum_{k \in R(X)} \left(\sum_{\omega \in A_k} \varphi(X(\omega))P(\omega) \right) = \sum_{k \in R(X)} \varphi(k) \cdot \mathbb{P}(A_k) = \sum_{k \in R(X)} \varphi(k) \cdot \mathbb{P}(X = k),$$

as desired. □

Example 5.12: Expectation of Flip and Roll

Let's revisit our coin flip/die roll example for the very last time. There, we had the sample space

$$\Omega = \{(c, d) : c = H \text{ or } T,\ d = 1, 2, \ldots, 6\}$$

and we assumed that all outcomes were equally likely, i.e., that $\mathbb{P}(\omega) = 1/12$ for each outcome. We considered the random variable

$$X(\omega) = X(c, d) = \begin{cases} 10 & c = H \text{ and } d \text{ is even} \\ -2 & c = H \text{ and } d \text{ is odd} \\ 1/2 & c = T \text{ and } d = 1, 2 \\ 0 & c = T \text{ and } d = 3, 4, 5, 6 \end{cases}$$

with range $R(X) = -2, 0, 1/2, 10$. By simply counting the number of outcomes in the events $\{X = k\}$ for $k \in R(X)$ and dividing by 12, we see immediately that

$$\mathbb{P}(X = k) = \begin{cases} 1/4 & k = -2 \\ 1/3 & k = 0 \\ 1/6 & k = 1/2 \\ 1/4 & k = 10 \end{cases}$$

for $k \in R(X)$. Thus, in view of Theorem 5.14, we have

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k \in -2, 0, 1/2, 10} k \cdot \mathbb{P}(X = k) \\ &= (-2)\mathbb{P}(X = -2) + (0)\mathbb{P}(X = 0) + (1/2)\mathbb{P}(X = 1/2) + (10)\mathbb{P}(X = 10) \\ &= \frac{-2}{4} + \frac{0}{3} + \frac{1/2}{6} + \frac{10}{4} \\ &= \frac{25}{12} \end{aligned}$$

which is precisely what we had found before (but with more effort).

Example 5.13: Expected Number of Flips

Consider an experiment in which we flip a coin n and keep track of the the outcome of each flip. This can be described by the sample space

$$\Omega = \{(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) : \epsilon_k = H \text{ or } T \text{ for } k = 1, 2, \ldots, n\}.$$

For example, if $n = 4$, $(H, T, H, T) = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ is the outcome in which heads came up on the first and third flips and tails on the second and forth. If we assume that the coin is biased, coming up heads on each flip with probability p and tails with probability $q = 1 - p$, and that each flip is independent of the others, we can compute the probability of the outcomes multiplicatively. For example,

$$\mathbb{P}(H, T, H, T) = p \cdot q \cdot p \cdot q = p^2 q^2.$$

Clearly, since the coin is biased, the $(\#(\Omega) = 2^n)$ outcomes in such an experiment are not equally likely and computing the probability of an outcomes comes down to how many heads vs. tails appear. Let's now introduce a random variable which keeps track of just that.

For this experiment, let X be the number of heads that appeared in the sequence of n flips. We note that $R(X) = \{0, 1, 2, \ldots, n\}$ for it is possible that no heads appear $X = 0$, all flips are heads, $X = n$ and that we can have every integer in between. Our aim is to compute the expectation of X. Using the definition, we have

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

You should convince yourself that you don't want to actually compute this summation. For example, when n is large, this summation has 2^n terms having values of X and probabilities which can vary by a lot. For this reason, let's instead attempt to use Theorem 5.14 to compute this expectation. To do this, we need to get a handle on the values of

$$\mathbb{P}(X = k) = \mathbb{P}(\text{In } n \text{ flips, exactly } k \text{ heads appear})$$

for then, we can compute the expectation via

$$\mathbb{E}(X) = \sum_{k \in R(X)} k\mathbb{P}(X = k).$$

For $k = 0 \in R(X) = \{0, 1, \ldots, n\}$,

$$\{X = k\} = \{X = 0\} = \{(T, T, \ldots, T)\}.$$

In other words, the event $\{X = k\}$ is the singleton event in which all flips resulted in tails. For this, we have

$$\mathbb{P}(X = 0) = \mathbb{P}(\{(T, T, \ldots, T\})) = q^n.$$

The event $\{X = 1\}$ is a little more complicated. This event consists of all of those outcomes in which H appears a single time and T appears in the other $n - 1$ flips. For example, when $n = 4$,

$$\{X = 1\} = \{(H, T, T, T), (T, H, T, T), (T, T, H, T), (T, T, T, H)\}.$$

In general, you should observe that there are exactly n outcomes in the event $\{X = 1\}$ – this is the number of choices you have to place a single H and then fill the rest with T's. Luckily, since each such outcomes has one H and $n - 1$ T's, each outcomes has probability pq^{n-1}. Putting these observations together, we have

$$\mathbb{P}(X = 1) = \#(\{X - 1\}) \times pq^{n-1} = npq^{n-1}.$$

Continuing this argument for a general $k \in \{0, 1, \ldots, n\}$, we see that the number of outcomes in $\{X = k\}$ is precisely the number of ways you can place k H's in n slots (and then fill the rest with T's). Thinking back to our exercises in counting, this is precisely

$$\#(\{X = k\}) = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Evidently, each such event has probability $p^k q^{n-k}$ and so we find that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k q^{n-k}$$

for $k \in \{0, 1, \ldots, n\}$. Thus, by Theorem 5.14, we have

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k}.$$

The only remaining question is how to compute this sum. Fortunately, there are several ways to do this. The way that we shall pursue here is called "the derivative trick" and, as we will see, it comes in handy in computations for several random variables.

To this end, let's forget for the moment that p and q are related and, in doing so, we can consider the more general summation

$$f(x) = \sum_{k=0}^{n} k \cdot \binom{n}{k} x^k q^{n-k} = \sum_{k=0}^{n} k x^k \binom{n}{k} q^{n-k}$$

which is precisely $\mathbb{E}(X)$ when evaluated at $x = p$. The big idea in the derivative trick is to recognize that kx^k almost looks like the power rule for the derivative of x^k and, as it turns out, this observation is a fruitful one. First, to align it exactly with the product rule, let's factor out a single power of x. We have

$$f(x) = \sum_{k=0}^{n} k x x^{k-1} \binom{n}{k} q^{n-k} = x \sum_{k=0}^{n} k x^{k-1} \binom{n}{k} q^{n-k}$$

and, upon noting that $\frac{d}{dx} x^k = k x^{k-1}$ for $k = 0, 1, \ldots, n$ and that the derivative is linear (meaning we can interchange it with scalar multiplication and summation),

$$f(x) = x \sum_{k=0}^{n} \left(\frac{d}{dx} x^k \right) \binom{n}{k} q^{n-k} = x \sum_{k=0}^{n} \frac{d}{dx} \left(\binom{n}{k} x^k q^{n-k} \right) = x \frac{d}{dx} \left(\sum_{k=0}^{n} \binom{n}{k} x^k q^{n-k} \right).$$

If we think back to the Chapter 3 and the binomial theorem, in particular, we recognize the term in parentheses can be written much more simply as $(x + q)^n$. Thus,

$$f(x) = x \frac{d}{dx} (x + q)^n = xn(x + q)^{n-1} = nx(x + q)^{n-1}$$

and, upon setting $x = p$, we have

$$\mathbb{E}(X) = f(p) = np(p + q)^{n-1} = np(1)^{n-1} = np.$$

Let us now treat two important corollaries of Theorem 5.14. The first gives us an equivalent way to calculate the variance and moments of a random variable in terms of the probabilities $\mathbb{P}(X = k)$ for $k \in R(X)$. The second applies directly to computing probabilities of events defined in terms of the random variable X.

Corollary 5.15. *Let Ω be a countable sample space with probability measure $\mathbb{P}$ and let X be a random variable on Ω with range $R(X)$ and mean $\mu = E(X)$. Providing the following series converge absolutely, we have*

$$\mathrm{Var}(X) = \sum_{k \in R(X)} (k - \mu)^2 \cdot \mathbb{P}(X = k) = \sum_{k \in R(X)} k^2 \cdot \mathbb{P}(X = k) - \mu^2$$

and, for each $n \in \mathbb{N}$,

$$\mathbb{E}(X^n) = \sum_{k \in R(X)} k^n \cdot \mathbb{P}(X = k).$$

Proof. The results follow immediately from Theorem 5.14 by applying (5.4) to $\varphi(x) = (x-\mu)^2$ for the variance and $\varphi(x) = x^n$ for the moments. □

Corollary 5.16. *Let* Ω *be a countable sample space with probability measure* $\mathbb{P}$ *and let* X *be a random variable on* Ω *with range* $R(X)$*. Then, for any subset* $I \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in I) = \sum_{k \in I \cap R(X)} \mathbb{P}(X = k).$$

Proof. Given $I \subseteq \mathbb{R}$, set $A = \{X \in I\}$ and observe that $\omega \in A$ if and only if $X(\omega) \in I$. From this it follows immediately that

$$\mathbb{1}_A(\omega) = \mathbb{1}_I(X(\omega))$$

for every $\omega \in \Omega$. Thus, by virtue of Proposition 5.12, we have

$$\mathbb{P}(X \in I) = \mathbb{P}(A) = \mathbb{E}(\mathbb{1}_A) = \mathbb{E}(\mathbb{1}_I(X))$$

Let us now make an appeal to Theorem 5.14. Specifically, using (5.4) with $\varphi(x) = \mathbb{1}_I(x)$, we have

$$\mathbb{E}(\mathbb{1}_I(X)) = \sum_{k \in R(X)} \mathbb{1}_I(k)\mathbb{P}(X = k) = \sum_{k \in I \cap R(X)} \mathbb{P}(X = k)$$

where we have used the fact that $\mathbb{1}_I(k) = 1$ if and only if $k \in I$. All together we have

$$\mathbb{P}(X \in I) = \mathbb{E}(\mathbb{1}_I(X)) = \sum_{k \in I \cap R(X)} \mathbb{P}(X = k)$$

as desired. □

We end this section by studying three different but surprisingly similar examples which demonstrate, minimally, the utility of Theorem 5.14 and its corollaries.

Example 5.14:

Consider the roll of a single perfect die which we describe via the sample space $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$ taken to be equipped with the probability measure $\mathbb{P}_1$ which is the uniform measure on Ω_1, i.e., $\mathbb{P}_1(A_1) = \#(A_1)/6$ for each event $A_1 \subseteq \Omega_1$. On Ω_1, consider the random variable

$$X_1(\omega) = \begin{cases} 0 & \omega = 1, 3 \\ -1 & \omega = 2, 6 \\ 1 & \omega = 4, 5 \end{cases}$$

for $\omega \in \Omega_1$. Observe that, $R(X_1) = \{-1, 0, 1\}$ and

$$\mathbb{P}_1(X_1 = k) = \frac{1}{3}$$

for $k = -1, 0, 1$. Using Theorem 5.14 and Corollary 5.15, we readily compute

$$\mathbb{E}(X_1) = \sum_{k \in R(X_1)} k \cdot \mathbb{P}_1(X_1 = k) = (-1)\frac{1}{3} + (0)\frac{1}{3} + (1)\frac{1}{3} = 0$$

and find similarly that $\text{Var}(X_1) = 2/3$. In fact, thanks to Corollary 5.16, we can compute the probability of any event associated to the random variable X_2 simply based on our knowledge of $\mathbb{P}_1(X_1 = k)$ for $k \in R(X_1)$.

Example 5.15:

Let's now consider the roll of a biased die which has the sample space $\Omega_2 = \{1,2,3,4,5,6\}$ but it is weighted so that even numbers are twice as likely as odd ones. In other words, we take Ω_2 to be equipped with the probability measure $\mathbb{P}_2$ with

$$\mathbb{P}_2(\omega) = \begin{cases} \frac{1}{9} & \omega = 1,3,5 \\ \frac{2}{9} & \omega = 2,4,6 \end{cases};$$

you should verify that this does define a probability measure $\mathbb{P}_2$ having the property that evens are twice as likely as odds. On Ω_2, consider the random variable

$$X_2(\omega) = \begin{cases} 0 & \omega = 1,2 \\ -1 & \omega = 3,4 \\ 1 & \omega = 5,6 \end{cases}$$

with range $R(X_2) = \{-1,0,1\}$. Observe that,

$$\mathbb{P}_2(X_2 = 0) = \mathbb{P}_2(\{1,2\}) = \mathbb{P}_2(1) + \mathbb{P}_2(2) = \frac{1}{9} + \frac{2}{9} = \frac{1}{3}.$$

Analogously, $\mathbb{P}_2(X_2 = -1) = \mathbb{P}_2(X_2 = 1) = 1/3$ so that

$$\mathbb{P}_2(X_2 = k) = \frac{1}{3}$$

for $k = -1,0,1$. Following from Theorem 5.14, we have $\mathbb{E}(X_2) = 0$ and $\text{Var}(X_2) = 2/3$ just like X_1. In fact, Corollary 5.16 shows that the probabilities for any event associated to X_2 is the same as it is associated to X_1.

Example 5.16:

Finally, let's consider the experiment of throwing a marble in a box of dimension 3×6. As we previously discussed, this can be represented as the sample space $\Omega_3 = \{(x,y) : 0 \leq x \leq 3,\ 0 \leq y \leq 6\} = [0,3] \times [0,6]$ and, provided that we take any location as equally likely, a reasonable probability measure is given by

$$\mathbb{P}_3(A_3) = \frac{\text{Area}(A_3)}{\text{Area}(\Omega)} = \frac{\text{Area}(A_3)}{18}$$

for any reasonable event/set $A_3 \subseteq \Omega_3$. Let's now consider a random variable X_3 which simply takes into account the x-value of the landing position which is defined by

$$X_3(\omega) = X_3(x,y) = \begin{cases} -1 & 0 \leq x < 1 \\ 0 & 1 \leq x < 2 \\ 1 & 2 \leq x \leq 3 \end{cases}$$

for $\omega = (x,y) \in \Omega_3$ and we note that $R(X_3) = \{-1,0,1\}$. Though this sample space Ω_3 is not countable we can still compute the values of $\mathbb{P}_3(X_3 = k)$ for $k \in R(X_3)$. First, since $X_3 = -1$ everywhere in the rectangle $[0,1) \times [0,6]$, we have

$$\mathbb{P}_3(X_3 = -1) = \frac{\text{Area}([0,1) \times [0,6])}{18} = \frac{1 \times 6}{18} = \frac{1}{3}.$$

Similarly, we find that

$$\mathbb{P}_3(X_3 = k) = \frac{1}{3}$$

for $k = -1, 0, 1$. Though we don't have a definition for $\mathbb{E}(X_3)$ on the uncountable sample space Ω_3, we can still calculate the right side of (5.3) and find that

$$\sum_{k \in R(X_3)} k \cdot \mathbb{P}_3(X_3 = k) = (-1)\frac{1}{3} + (0)\frac{1}{3} + (1)\frac{1}{3} = 0.$$

In view of Theorem 5.14, it is reasonable to interpret $\mathbb{E}(X_3) = 0$. By a similar computation, we use Corollary 5.15 to make the interpretation that $\text{Var}(X_3) = 2/3$.

Let's make two very important observations about the three preceding examples. First, though the sample spaces $\Omega_1, \Omega_2, \Omega_3$, probability measures $\mathbb{P}_1$, $\mathbb{P}_2, \mathbb{P}_3$ and random variables X_1, X_2, X_3 were all different, the ranges of the random variables X_1, X_2 and X_3 were exactly the same $\{-1, 0, 1\}$ and we found that

$$\mathbb{P}_j(X_j = k) = \frac{1}{3}$$

for $j = 1, 2, 3$ and $k = -1, 0, 1$. Consequently, these random variables all have the same mean and variance and the probabilities associated to each one are exactly the same. If you think of these random variables as measurements and you could compute the probability (say, via, frequency) and along with it expectations, variances and moments – all of which are identical for each experiment – could you tell which experiment you were looking at? Think of this as a black box where all you could measure was the value of the random variable, could you tell which experiment was going on inside the black box? I think your answer is very likely "no". In fact, if one is only able to measure the values of a random variable X, the specific underlying experiment/sample space Ω and the specific nature of the probability measure $\mathbb{P}$ cannot generally be understood through only the knowledge of these values as these examples show. But, seen, through Theorem 5.14 and its corollaries, everything about that random variable (or probabilistic interest) can be answered by simply knowing its range and the probability that it takes the values in its range. This is the great advantage of Theorem 5.14: we are able to suppress just about every aspect of the underlying experiment and focus simply on the things associated to the random variable – they are all that matter. This is precisely the tack we take in the following subsections.

Our second observation involves the third example above. As we noted, the underlying sample space Ω_3 was uncountable (so that we do not, a prior, have a definition for expectation), yet everything still made sense in view of Theorem 5.14 and its corollaries. In fact, if one thinks carefully through our presentation, for the right hand sides of (5.3) and (5.4) to make sense the only thing that is really needed is the countability of $R(X)$ not the countability of Ω itself. This leads us smoothly into the next subsection.

5.2 Discrete Random Variables

In light of our previous discussion, we make the following definition which, notably, does not require the underlying sample space to be countable.

Definition 5.17 (Discrete Random Variables and the Probability Mass Function)**.** *A random variables X (on a sample space Ω equipped with probability $\mathbb{P}$) is said to be discrete if its range $R(X)$ is a countable set of real numbers. For such a random variable, we define the probability mass function p_X associated to X by*

$$p_X(k) = \mathbb{P}(X = k)$$

for all $k \in \mathbb{R}$.

Remark 5.18. Though probability mass functions are defined for every real number k, they can only be non-zero for those $k \in R(X)$. To see this, suppose that $k' \notin R(X)$ and so $\{X = k'\} = \varnothing$. Then $p_X(k') = \mathbb{P}(X = k') = \mathbb{P}(\varnothing) = 0$. For this reason, we will often only apply $p_X(k)$ to $k \in R(X)$ and this can also be considered its domain of definition.

As we observed throughout the last section, a random variable on a countable sample space Ω necessarily has countable range and so is discrete. Of course, the last example of the previous subsection shows that discrete random variables also exist on uncountable sample spaces. Though discrete random variables are not the only random variables, as we will see, they will be the focus of this section. Lets first focus our attention on probability mass functions. First, by the axioms of probability, we have that $p_X(k) = \mathbb{P}(X = k) \geq 0$ and, as we pointed out, this can only be positive for $k \in R(X)$. Furthermore, by the additivity of probability, we have

$$\sum_{k \in R(X)} p_X(k) = \sum_{k \in R(X)} \mathbb{P}(X = k) = \mathbb{P}\left(\bigcup_{k \in R(X)} \{X = k\} \right) = \mathbb{P}(\Omega) = 1$$

where we have used the fact that $\{X = k\}$ indexed by $k \in R(X)$ is a partition of the sample space Ω. Putting these observations together, we have established the proof of necessity in the following proposition; we leave the proof of sufficiency to the curious reader.

Proposition 5.19 (A characterization of probability mass functions)**.** *The probability mass function p_X of a discrete random variable X satisfies the following properties:*

- *$p_X(k) \geq 0$ for all $k \in \mathbb{R}$ and, further, $p_X(k) > 0$ only for $k \in R(X)$.*
- *We have*
$$\sum_{k \in R(X)} p_X(k) = 1.$$

Conversely, any function $p : \mathbb{R} \to \mathbb{R}$ which satisfies the above two properties which is non-zero only on some countable set R is the probability mass function of some random variable X, i.e., $p = p_X$, and $R = R(X)$.

Example 5.17: A Simple PMF

As we observed, the final three thee examples in the preceding subsection all have the same probability mass function

$$p(k) = \frac{1}{3}$$

for $k = -1, 0, 1$ and 0 otherwise. Observe that this function is non-negative, non-zero only on the countable set $R(X) = \{-1, 0, 1\}$ and it has

$$\sum_{k \in R} p(k) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1.$$

As $p = p_{X_1} = p_{X_2} = p_{X_3}$, the above illustrates that several random variables can have the same probability mass function. Further, if all we are concerned with is the value of these random variables (independently from anything else) and their associated probabilities, this probability mass function tells us everything we'd need to know.

Example 5.18: Waiting for Heads

We recall the experiment discussed in Example 5.7 – need to reference other appearance as well wherein we flipped a biased coin ($0 \leq p < 1$) over and over until heads appears. As we computed, the random variable N, which counts the number of flips until heads appears, has $R(N) = \{1, 2, \dots\} = \mathbb{N}_+$ and

$$p_N(k) = \mathbb{P}(N = k) = q^{k-1}p$$

for all $k \in \mathbb{N}_+$ and 0 otherwise. As q and p are non-negative, p_N is a non-negative function which is non-zero

only on the countable set $R(N) = \mathbb{N}_+$. Further, using the formula for the sum of a geometric series,

$$\sum_{k \in R(N)} p_N(k) = \sum_{k=1}^{\infty} q^{k-1} p = p\left(\sum_{k=1}^{\infty} q^{k-1}\right) = p\frac{1}{1-q} = \frac{p}{p} = 1$$

as necessary.

Example 5.19: PMF of Poisson

Given $\lambda > 0$, consider $p(k)$ defined by

$$p(k) = C\frac{\lambda^k}{k!}$$

for $k = 0, 1, 2 \dots$ and $p(k) = 0$ otherwise. We ask: If $p(k)$ is the probability mass function of a random variable X, i.e., $p = p_X$, what is the value of C?
To answer this, we first observe that $p(k)$ must be non-negative and so we require that $C \geq 0$. Further, we must have

$$1 = \sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} C\frac{\lambda^k}{k!} = C\left(\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\right).$$

Now, the series in the parentheses above we recall from our calculus class. This is precisely the series for e^λ with this as its sum. Therefore $1 = Ce^\lambda$ and so $C = 1/e^\lambda = e^{-\lambda}$. In view of the preceding proposition, we may conclude that

$$p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for $k \in \mathbb{N} = \{0, 1, 2, \dots\}$ is a probability mass function of a random variable X with $R(X) = \mathbb{N}$. Any such random variable, as we will soon discuss, is called a **Poisson random variable with parameter** λ in honor of the mathematician and physicist Siméon Denis Poisson.

We are now ready to discuss the expectation of discrete random variables stated in terms of the probability mass function.

Theorem 5.20. *Let X be a discrete random variable (on a sample space Ω equipped with probability $\mathbb{P}$) with range $R(X)$ and probability mass function p_X. Then:*

1. *We have*

$$\mathbb{E}(X) = \sum_{k \in R(X)} k \cdot p_X(k)$$

provided this series converges absolutely.

2. *For any function $\varphi : \mathbb{R}$,*

$$\mathbb{E}(\varphi(X)) = \sum_{k \in R(X)} \varphi(k) p_X(k)$$

provided this series converges absolutely.

3. *If X has mean $\mu = \mathbb{E}(X)$, then*

$$\mathrm{Var}(X) = \sum_{k \in R(X)} (k-\mu)^2 p_X(k) = \left(\sum_{k \in R(X)} k^2 p_X(k)\right) - \mu^2$$

which is said to be finite when the above series converge and infinite otherwise.

4. For a natural number n,

$$\mathbb{E}(X^n) = \sum_{k \in R(X)} k^n p_X(k)$$

provided this converges absolutely. For any subset $I \subseteq \mathbb{R}$,

$$\mathbb{P}(X \in I) = \sum_{k \in I \cap R(X)} p_X(k).$$

Since $p_X(k) = \mathbb{P}(X = k)$, this theorem follows immediately from Theorem 5.14 and its corollaries when Ω is a countable sample space. The proof in the case that Ω is not countable (and the existence of expectation in that case) is somewhat beyond the scope of this course and you will see it if you take a graduate course in probability.

Throughout all of our discussion, we have concluded that all questions (of probabilistic interest) about discrete random variables can be answered in terms of their probability mass functions. Thus, we shall take the opportunity to introduce several important classes of discrete random variables by introducing them via their probability mass functions. As we know random variables with the same probability mass functions can arise on many different sample spaces, we will discuss the types of experiments that give rise to these random variables, i.e., we describe certain experiments for which these are good models. We treat these systematically throughout the following four subsections.

5.2.1 Bernoulli Random Variables

Definition 5.21. *Let $0 \leq p \leq 1$ be a fixed number and set $q = 1 - p$. A Bernoulli random variable with parameter p is a random variable X which has probability mass function*

$$p_X(k) = \begin{cases} p & k = 1 \\ q & k = 0 \end{cases}$$

and 0 otherwise. In this case, we shall write $X \sim \text{Ber}(p)$.

As we saw in Example 5, Bernoulli random variables can arise in experiments where one flips a single biased coin and assigns the value 1 if heads and 0 if tails. More generally, Bernoulli random variables represent experiments where there are two principal outcomes, "success" and "failure" and the associated random variable assigns the value 1 to success and 0 to failure. Though this random variable might appear somewhat uninteresting, it is the basis for the study of much more complicated and interesting random variables, one case of which is introduced in the following subsection.

For $X \sim \text{Ber}(p)$, we compute

$$\mathbb{E}(X^n) = (1)^n p + (0)^n q = p$$

for all n an so, in particular, we have

$$\mathbb{E}(X) = p \qquad \text{and} \qquad \text{Var}(X) = pq$$

because $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p - p^2 = p(1-p) = pq$.

5.2.2 The Binomial Random Variable

Definition 5.22. *Let $0 \leq p \leq 1$ be fixed, set $q = 1 - p$ and let n be a positive natural number. A Binomial random variable with parameters p and n is a random variable X with probability mass function*

$$p_X(k) = \binom{n}{k} p^k q^{n-k}$$

for $k = 0, 1, 2, \ldots, n$ and 0 otherwise. In this case, we shall write $X \sim \text{Bin}(p, n)$

As we found, binomial random variables represent the number of heads that appear in n consecutive and independent flips of a biased coin. More generally, it represents any random variable that counts the number of successes (which happen with probability p) in n independent Bernoulli trials. We previously computed

$$\mathbb{E}(X) = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k} = np$$

using the so-called derivative trick. Let's use the derivative trick once again to compute the second moment for the binomial random variable. First, upon noting that $X^2 = X(X-1) + X$, the linearity of expectation gives

$$\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X) = \mathbb{E}(X(X-1)) + np$$

and this reduces out problem of computing the expectation of $X(X-1)$ which, as we will see, is slightly easier to do. We have

$$\mathbb{E}(X(X-1)) = \sum_{k=0}^{n} k(k-1)\binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} k(k-1) p^k \binom{n}{k} q^{n-k} = g(p)$$

where

$$g(x) = \sum_{k=0}^{n} k(k-1) x^k \binom{n}{k} q^{n-k}$$

With the aim as recognizing a second derivative in the terms $k(k-1)x^k$, we factor x^2 out of the sum to find that

$$g(x) = x^2 \sum_{k=0}^{n} k(k-1) x^{k-2} \binom{n}{k} q^{n-k} = x^2 \sum_{k=0}^{n} \frac{d^2}{dx^2}\left(x^k\right) \binom{n}{k} q^{n-k}$$

where we have noted that $\frac{d^2}{x^k} = k(k-1)x^{k-2}$. Using the linearity of the second derivative and appealing to the binomial theorem, we find

$$g(x) = x^2 \sum_{k=0}^{n} \frac{d^2}{dx^2}\left(\binom{n}{k} x^k q^{n-k}\right) = x^2 \frac{d^2}{dx^2}\left(\sum_{k=0}^{n} \binom{n}{k} x^k q^{n-k}\right) = x^2 \frac{d^2}{dx^2}(x+q)^n$$

or, equivalently,

$$g(x) = n(n-1)x^2(x+q)^{n-2}.$$

When $x = p$, this gives

$$\mathbb{E}(X(X-1)) = g(p) = n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2(1)^{n-2} = n(n-1)p^2.$$

Consequently,

$$\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + np = n(n-1)p^2 + np = np - np^2 + n^2p^2.$$

Using our formula for the variance in terms of the mean and the second moment, this gives

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = np - np^2 + n^2p^2 - (np)^2 = np - np^2 = np(1-p) = npq.$$

We have established the following.

Proposition 5.23. *Given, $n \in \mathbb{N}_+$ and $0 \leq p \leq 1$, the binomial random variable $X \sim Bin(n,p)$ has*

$$\mathbb{E}(X) = np \qquad \text{and} \qquad \mathrm{Var}(X) = npq$$

where $q = 1 - p$.

As we previously remarked, the Bernoulli random variable is a special case of the binomial random variable in the case that $n = 1$. In this case, we note that the above proposition recaptures the mean and variance for the Bernoulli random variable. Further, with the interpretation of the mean and variance as measuring the center of mass (balance point) and the spread, it is interesting to note that both scale proportionally to n. Perhaps this remark should be made in terms of σ. This is illustrated in the Figure 5.3

The following exercise gives a nice variation of the Binomial random variable.

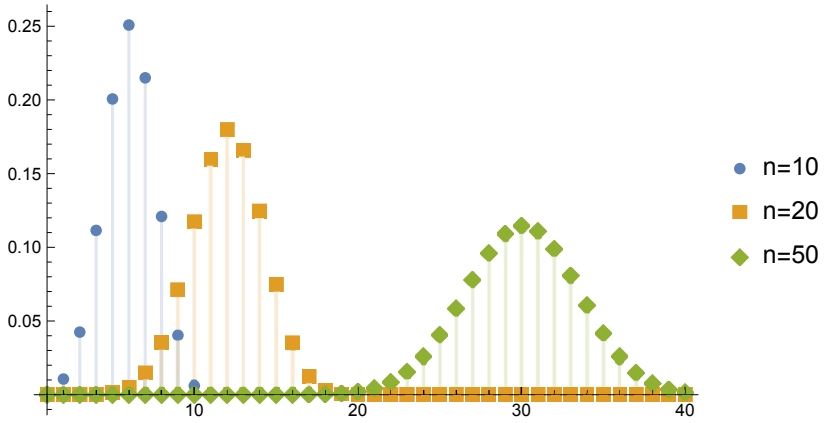


Figure 5.3: The PMF of Binomial random variables with $p = \frac{3}{5}$ and $n = 10, 20, 50$

Exercise 5.6: Alice's Random Walk Revisited

In Exercise 4.2.1, we studied Alice's random walk and determined that Alice's position after n steps was described using combination coefficients. In this problem, we will change the rules of Alice's game (only slightly) and study her position S_n from the perspective of random variables. For this variation of the game, at each step, Alice will flip a biased coin with probability p of coming up heads and $q = 1 - p$ of coming up tails. As before, Alice will start at 0 and walk according to the same rules ($+1$ for heads and -1 for tails) flipping the (now biased) coin independently at each step.
For each n, we can describe Alice's sample space as

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots, \omega_n) : \omega_k = \text{“heads” or “tails” for each step } 1 \leq k \leq n.\}$$

1. Find a collection of random variables $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ for which

   $$S_n = X_1 + X_2 + \cdots + X_n.$$

   and, for each, describe its probability mass function and find its expectation. That is, compute

   $$p_{X_j}(k) = \mathbb{P}(X_j = k)$$

   for each $j = 1, 2, \ldots, n$ and $k \in \mathbb{R}$ and also compute $\mathbb{E}(X_j)$ for each $j = 1, 2, \ldots, n$. What do the random variables $X_1, X_2, \ldots, X_n$ represent?

2. Using the ideas of the previous homework, compute

   $$p_n(k) = \mathbb{P}(S_n = k).$$

3. Use your knowledge of the previous item to compute Alice's expected position, $\mathbb{E}(S_n)$. When $p = q = 1/2$, explain why your answer makes sense.

4. Since $S_n = X_1 + X_2 + \cdots + X_n$, confirm your answer using the identity

   $$\mathbb{E}(S_n) = \mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$$

   (which is valid in view of the previous problem).

Another exercise here.

5.2.3 Geometric Random Variable

Definition 5.24. *Let $0 \leq p \leq 1$ be fixed and set $q = 1 - p$. A random variable X is said to be geometric with parameter p if it is a random variable with probability mass function*

$$p_X(k) = pq^{k-1}$$

for $k \in \mathbb{N}_+ = \{1, 2, \ldots, \}$. Here, we will write $X \sim \text{Geo}(p)$.

Looking back to Example 5.18, the random variable N which counts the number of independent (biased) coin flips until the first appearance of heads is a geometric random variable. Generally speaking, geometric random variables are those that measure "to first instance" in independent trials. As you showed in Exercise 5.4, we have the following:

Proposition 5.25. *Let $X \sim \text{Geo}(p)$. Then*

$$\mathbb{E}(X) = \frac{1}{p} \qquad \textit{and} \qquad \text{Var}(X) = \frac{1-p}{p^2}.$$

Exercise 5.7: Mike's Backward Craps

Michael walks into a casino and considers playing the following game. Two perfect dice are rolled over and over again (and independently) until the sum of their faces is 7 or 11. Let N be the number of rolls until 11 or 7 is observed for the first time. We shall denote Michael's winnings by M. If $N = 1$, then Michael has to pay 80 dollars, i.e., $M = -80$. If $N = 2$, then Michael wins 50 dollars, i.e., $M = 50$. If $N > 2$, then Michael wins $M = 5(N - 1)$ dollars. Note that Michael wins money if he makes it past the first roll!

1. Compute the probability mass function p_M.
2. Compute the expected value $\mathbb{E}(M)$ of Michael's winnings.
3. Should Michael play this game?

Exercise 5.8: Using a biased coin to simulate a fair one

Suppose that you would like to simulate a fair coin flip, but only have a biased coin which comes up heads with probability p for some $0 < p < 1$. Consider the following procedure:

i. Flip the coin.

ii. Flip the coin again.

iii. If both flips land on heads or both land on tails, return to Step 1.

Let the result of the last coin flip be the result of the experiment.

1. Show that the probability that this experiment results in heads is $1/2$. Hint: For each $k = 1, 2, \ldots$, let A_k be the event that the experiment yields heads after $2k$ coin flips, i.e., after flipping the coin twice k times.
2. Compute the probability that this experiment results in tails. Note: You may appeal to the calculations done in the previous part, if helpful, while denoting the analogous relevant events by B_k for $k = 1, 2, \ldots$.
3. Does this experiment actually simulate flipping a fair coin?

4. Consider the random variable N which counts the number of coin flips until the deciding coin flip of the experiment. For example, if the first two coin flips resulted in $\{T, H\}$ (so that Steps 1 and 2 didn't need repeating), the experiment was decided to be H after the two flips and so $N = 2$. What is the range of the random variable N?
5. Compute the probability mass function $p = p_N$ of N. Hint: Observe that $\{H = 2\} = A_1 \cup B_1$.
6. Compute the expected value of N, i.e, compute $E[N]$.

5.2.4 Poisson Random variable

The final discrete random variable we will discuss is the so-called Poisson random variable which we briefly introduced in Example 5.19.

Definition 5.26. *Let $\lambda > 0$. A discrete random variable X is said to be Poisson with parameter λ if its probability mass function is given by*

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for $k \in \mathbb{N} = \{0, 1, 2, \dots\}$. In this case, we shall write $X \sim \text{Pois}(\lambda)$.

In contrast to Bernoulli, Binomial, and Geometric random variables (or the position of Alice's random walk), Poisson random variables are a little bit harder to describe from first principles. Generally speaking, Poisson random variables are used to model the number of events that occur in a given time interval where the events happen independently and each is rare in the sense that it happens with low probability in any small window of time. For example, Poisson random variable might model the number of customers that arrive at a post office in a given interval of time and the corresponding λ will be proportional to the length of the time interval. Poisson random variables are also commonly used to number of radioactive isotopes that decay in a given length of time $[0, T]$ where λ is proportional to T.

In what follows, we show how the Poisson random variable arises naturally as the limit of binomial random variables. Later in this subsection, we shall see how Poisson random variables arise as random variables which count the number of rare and independent occurrences in continuous time intervals.

Example 5.20: Derivation of the Poisson as a limit of Binomial

Let's consider a situation in which $X \sim \text{Bin}(n, p)$ is a binomial random variable where we assume that n is extremely large and p is extremely small in such a way that their product is some (not too small or too large) number $\lambda = np$. For example, X could represent that number of occurrences of an extremely rare event (with probability $p \ll 1$) in a very large number ($n \gg 1$) of trials. To compute the probability that there are k successes we wish to then estimate

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

for n large and p small such that $\lambda \approx np$. If we write $X = X_n$ to indicate X's dependence on n so that $X_n \sim \text{Bin}(n, \lambda/n)$, what we're looking to understand is the probability mass function of the (necessarily discrete) random variable

$$Y = \lim_{n\to\infty} X_n$$

which is[a]

$$p_Y(k) = \lim_{n\to\infty} p_{X_n}(k) = \lim_{n\to\infty} \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

for each $k \in \mathbb{N}$. Here, we note that for each fixed $k \in \mathbb{N}$, n is eventually greater than k and so $p_{X_n}(k)$ is

given by the above expression. Observe that

$$\begin{aligned}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} &= \frac{n!}{(n-k)!k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\\ &= \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\end{aligned}$$

for each $k \leq n$. Since the limit of the product is the product of limits (when they exist), let's compute three limits that will allow us to compute $p_Y(k)$. First,

$$\begin{aligned}\lim_{n\to\infty}\frac{n!}{(n-k)!n^k} &= \lim_{n\to\infty}\frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k}\\ &= \lim_{n\to\infty}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)\\ &= 1\cdot 1\cdot 1\cdots 1\\ &= 1.\end{aligned}$$

Second, by virtue of Proposition A.1 (using $L=-\lambda$), we find that

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Finally, we observe that since k is fixed, the continuity of the power function gives

$$\begin{aligned}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k} &= \left(\lim_{n\to\infty}\left(1-\frac{\lambda}{k}\right)\right)^{-k}\\ &= (1-0)^{-k}\\ &= 1.\end{aligned}$$

Putting these three limits together gives

$$\begin{aligned}p_Y(k) &= \lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\ &= \frac{\lambda^k}{k!}\lim_{n\to\infty}\frac{n!}{(n-k)!n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}\\ &= \frac{\lambda^k}{k!}\cdot 1\cdot e^{-\lambda}\cdot 1\\ &= e^{-\lambda}\frac{\lambda^k}{k!}\end{aligned}$$

In this way, we see (in a way we will later make precise) that the limit of binomial random variables $X_n \sim \text{Bin}(n, \lambda/n)$ is a Poisson random variable Y with parameter λ.

[a] In the final chapter of these notes, we will discuss the convergence of random variables properly. Here, we are using (albeit formally) a version of convergence of random variables which is called *convergence in distribution*.

Figures 5.4 and 5.5 illustrate the probability mass function and cumulative distribution functions, respectively, of Poisson random variables for the values of $\lambda = 10$, 20, and 30. In line with the argument made in the preceding example, it is noteworthy that the probability mass function of the binomial random variables in Figure 5.3 resembles those of the Poisson random variables in Figure 5.4, especially so as n grows.

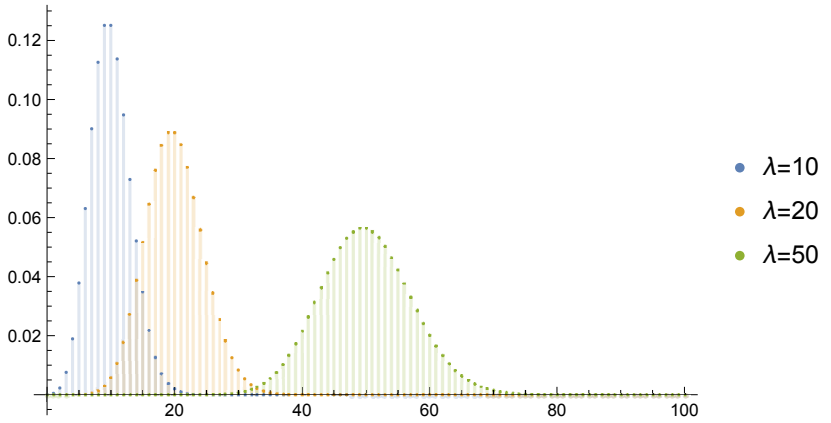


Figure 5.4: The PMF of Poisson Random Variables with $\lambda = 10$, 20 and 30.

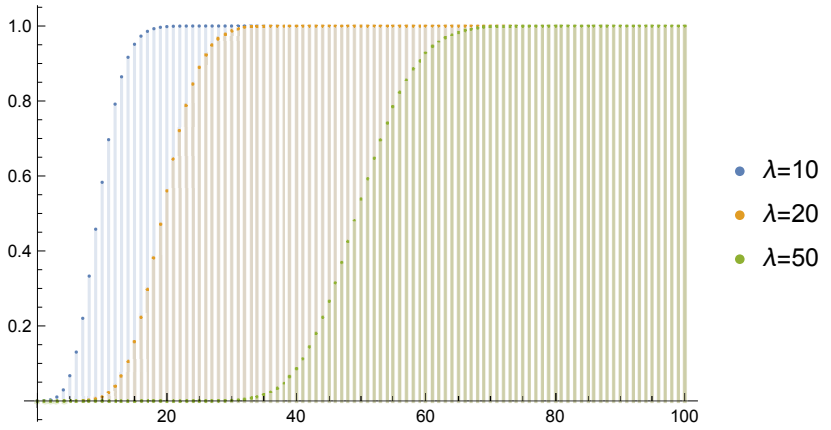


Figure 5.5: The CDF of Poisson Random Variables with $\lambda = 10$, 20 and 30.

Proposition 5.27. *Let $X \sim \text{Pois}(\lambda)$ for $\lambda > 0$. Then*

$$\mathbb{E}(X) = \lambda \qquad \textit{and} \qquad \text{Var}(X) = \lambda.$$

Proof. Here, we will prove that $\mathbb{E}(X) = \lambda$ and leave the computation of the variance as an exercise. Upon noting that $R(X) = \{0, 1, \dots\} = \mathbb{N}$, we appeal to Theorem 5.41 so see that

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!}. \tag{5.5}$$

To compute this series, let's argue using the "derivative trick". To this end, we see that

$$\begin{aligned}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} &= \lambda\sum_{k=0} k\frac{\lambda^{k-1}}{k!}\\ &= \lambda\sum_{k=0}^{\infty}\frac{d}{dx}\left(\frac{x^k}{k!}\right)\Big|_{x=\lambda}\\ &= \lambda\frac{d}{dx}\left(\sum_{k=0}^{\infty}\frac{x^k}{k!}\right)\Big|_{x=\lambda}\\ &= \lambda\frac{d}{dx}\left(e^x\right)\Big|_{x=\lambda}\\ &= \lambda e^{\lambda}\end{aligned}$$

where have noted that $\frac{d}{dx}x^k = kx^{k-1}$ for $k = 0, 1, \ldots$ and invoked the property that the derivative of the series for e^x can be taken term by term [15, Theorem 8.1]. Though I have used the derivative trick in the above computation, it should be noted that there is an easier way to obtain the identity above. Indeed, upon recognizing that the series on the left-hand side has 0 for its zeroth term and $k/k! = 1/(k-1)!$ for $k \geq 1$, we have

$$\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}.$$

By reindexing this series by $j = k-1$, we see that

$$\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = e^{\lambda}$$

so that

$$\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = \lambda\sum_{k=1}^{\infty}\frac{\lambda^{(k-1)}}{(k-1)!} = \lambda e^{\lambda}$$

as we had previously seen. With this identity, (5.5) gives

$$\mathbb{E}(X) = e^{-\lambda}\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

as was asserted. □

Exercise 5.9: Moments of Poisson

Let $X \sim \text{Pois}(\lambda)$. Show that, for each n,

$$\mathbb{E}(X^n) = \lambda\mathbb{E}((X+1)^{n-1}).$$

Since $\mathbb{E}(X) = \lambda$, use this to show that $\text{Var}(X) = \lambda$ and also compute the moments $\mathbb{E}(X^2)$ and $\mathbb{E}(X^3)$.

One advantage of knowing the the mean of a Poisson random variable with parameter λ is itself λ, we can compute many probabilities by simply knowing the mean.

Example 5.21: Customers arrive at a bank

A bank in rural Maine opens at 9:00AM on weekdays. On any given day, that bank expects two customers to arrive in the first hour. In fact, after a long-term study, the bank finds that the number of customers to arrive within the first hour can be modeled by a Poisson random variable N. On any given day, what is the probability that no customers will arrive in the first hour?

To compute this probability, we first note that the information given gives us that N is a Poisson random variable with $\mathbb{E}(N) = 2$ and hence, in view of the previous proposition, $N \sim \text{Pois}(2)$. Thus,

$$\mathbb{P}(\text{No customers arrive in the first hour}) = p_N(0) = e^{-2}\frac{2^0}{0!} = e^{-2} \approx 0.135.$$

We conclude this subsection with a derivation of the Poisson random variable from three basic assumptions. In the derivation, we readily make use of "little O" notation which is a common notation used in mathematics and its applications; some background on this common mathematical notation can be found in the appendix.

Example 5.22: A Derivation of Poisson Random Variable

As we discussed, Poisson random variables can be used to model the number of occurrences of rare and independent events in a given time period. In what follows, we shall make three hypotheses concerning the probability of occurrences of some event in a time interval[a] $I = (0, T]$ and use them to derive a random a Poisson random variable X measuring the number of occurrences in the interval I. We make the following assumptions:

There is a positive number λ and two real-valued functions α and β with $\alpha(t) = o(t)$ and $\beta(t) = o(t)$ as $t \to 0$ such that:

1. Given any collection of disjoint (non-overlapping) subintervals $I_1, I_2, \ldots, I_L$ of I, the collection of events $E_1, E_2, \ldots, E_L$, defined by
$$E_l = \{\text{There is an occurence in } I_l\}$$
for $l = 1, 2, \ldots, L$, is an independent collection.

2. For any (subinterval) $J \subseteq I$ of length t,
$$\mathbb{P}(\{\text{A single occurrence is observed in } J\}) = \lambda t + \alpha(t).$$

3. For any subinterval $J \subseteq I$ of length t,
$$\mathbb{P}(\{\text{At least two occurrences are observed in } J\}) \leq \beta(t).$$

Under these three hypotheses, we will show that

$$X := \#(\text{Occurrences in } I) \sim \text{Pois}(\lambda T).$$

by showing that, for each $k \in \mathbb{N}$,

$$\mathbb{P}(X = k) = e^{-\lambda T}\frac{(\lambda T)^k}{k!}.$$

To compute the probability of $\{X = k\}$ for $k \in \mathbb{N}$, it is useful to first partition I into n regular disjoint subintervals of time $I_1, I_2, \ldots, I_n \subseteq I$ each defined by

$$I_j = \left(T\frac{(j-1)}{n}, T\frac{j}{n}\right]$$

for $j = 1, 2, \ldots, n$. With these subintervals in hand, we can make use of our hypotheses to compute the probability of occurrences happening (or not happening) within time intervals in terms of α and β. We will then allow these subintervals to become small thereby isolating the subintervals of time in which occurrences happen individually, i.e., by letting $n \to \infty$, and this will give us precisely the probability of the event $\{X = k\}$.

With our strategy outlined, let's express $\{X = k\}$ as the disjoint union

$$\{X = k\} = A \cup B$$

where

$$A = \{\text{There are } k \text{ subintervals in which exactly one event occurs and } n - k \text{ in which no event occurs}\}$$

and

$$B = \{\text{There are } k \text{ occurences and at least one subinterval contains 2 (or more) occurrences}\}.$$

We first focus our attention on B. Observe that

$$B \subseteq \bigcup_{j=1}^{n} B_j$$

where

$$B_j = \{\text{The subinterval } I_j \text{ contains two or more occurrences}\}$$

for $j = 1, 2, \ldots, n$. Because the interval I_j has length T/n, the third hypothesis guarantees that

$$\mathbb{P}(B_j) = \beta\left(\frac{T}{n}\right)$$

$j = 1, 2, \ldots, n$. By virtue of the union bound (Theorem 2.11) and the monotonicity of probability, we have

$$0 \leq \mathbb{P}(B) \leq \sum_{j=1}^{n} \mathbb{P}(Bj) \leq \sum_{j=1}^{n} \beta\left(\frac{T}{n}\right) = n\beta\left(\frac{T}{n}\right).$$

Using the fact that $\beta(t) = o(t)$ as $t \to 0$, observe that

$$\lim_{n\to\infty} n\beta\left(\frac{T}{n}\right) = \lim_{t\to 0} T\frac{\beta(t)}{t} = T\lim_{t\to 0}\frac{\beta(t)}{t} = 0$$

and so, by virtue of the preceding estimate and the squeeze theorem, we conclude that

$$\lim_{n\to\infty} \mathbb{P}(B) = 0. \tag{5.6}$$

With this observation in hand, let's turn our attention to the event A. Using the first and second hypotheses, we find ourselves in the setting of Bernoulli trials: An occurrence happens in k subintervals (success) each with probability

$$p = \lambda\left(\frac{T}{n}\right) + \alpha\left(\frac{T}{n}\right)$$

and no occurrence happens in $n - k$ subintervals each with probability

$$\begin{aligned} q &= 1 - \lambda\left(\frac{T}{n}\right) - \alpha\left(\frac{T}{n}\right) - \beta\left(\frac{T}{n}\right) \\ &= 1 - \lambda\left(\frac{T}{n}\right) - \gamma\left(\frac{T}{n}\right) \end{aligned}$$

where $\gamma := \alpha + \beta$ and these successes and failures happen independently in view of the first hypothesis. We remark that it isn't quite true that $q = 1 - p$ because having no occurrences in the interval isn't the complement of having one occurrence. In any case, we have

$$\mathbb{P}(A) = \binom{n}{k} p^k q^{n-k} = \binom{n}{k} \left(\lambda\left(\frac{T}{n}\right) + \alpha\left(\frac{T}{n}\right)\right)^k \left(1 - \lambda\left(\frac{T}{n}\right) - \gamma\left(\frac{T}{n}\right)\right)^{n-k}$$

Now,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(A) &= \lim_{n\to\infty} \binom{n}{k} \left(\lambda\left(\frac{T}{n}\right) + \alpha\left(\frac{T}{n}\right)\right)^k \left(1 - \lambda\left(\frac{T}{n}\right) - \gamma\left(\frac{T}{n}\right)\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \frac{(\lambda T)^k}{n^k} \left(1 + \frac{1}{\lambda}\frac{n}{T}\alpha\left(\frac{T}{n}\right)\right)^k \left(1 + \frac{1}{n}\left(-\lambda T - n\gamma\left(\frac{T}{n}\right)\right)\right)^{n-k} \\
&= \frac{(\lambda T)^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!n^k} \left(1 + \frac{1}{\lambda}\frac{n}{T}\alpha\left(\frac{T}{n}\right)\right)^k \left(1 + \frac{a_n}{n}\right)^{n-k}
\end{aligned}$$

where

$$a_n = -\lambda T - n\gamma\left(\frac{T}{n}\right)$$

As we saw in Example 5.20,

$$\lim_{n\to\infty} \frac{n!}{(n-k)!n^k} = 1. \tag{5.7}$$

Given that $\alpha(t) = o(t)$ as $t \to 0$, we see that

$$\lim_{n\to\infty} \frac{n}{T}\alpha\left(\frac{T}{n}\right) = \lim_{t\to 0} \frac{\alpha(t)}{t} = 0$$

and so, by virtue of the continuity of $u \mapsto u^k$,

$$\lim_{n\to\infty} \left(1 + \frac{1}{\lambda}\frac{n}{T}\alpha\left(\frac{T}{n}\right)\right)^k = \left(1 + \frac{1}{\lambda}\lim_{n\to\infty} \frac{n}{T}\alpha\left(\frac{T}{n}\right)\right)^k = (1+0)^k = 1. \tag{5.8}$$

By a similar argument, we see that

$$\lim_{n\to\infty} a_n = -\lambda T - \lim_{n\to\infty} n\gamma\left(\frac{T}{n}\right) = -\lambda T - 0 = -\lambda T$$

where we have used the fact that $\gamma(t) = \alpha(t) + \beta(t) = o(t)$ as $t \to 0$. With this,

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^{-k} = (1+0)^{-k} = 1$$

and so, by an appeal to Proposition A.1, we have

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^{n-k} = \lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^{n} = e^{-\lambda T}. \tag{5.9}$$

Upon combining (5.7), (5.8), and (5.9), we obtain

$$\lim_{n\to\infty} \mathbb{P}(A) = \frac{(\lambda T)^k}{k!} 1 \cdot 1 \cdot e^{-\lambda T} = e^{-\lambda T} \frac{(\lambda T)^k}{k!}.$$

As we previously discussed, our hypotheses (written in terms of α and β) allow us to compute the probability of $\{X = k\}$ only when the size of the subintervals becomes arbitrarily small, i.e, as $n \to \infty$. Precisely, we have

$$\mathbb{P}(X = k) = \lim_{n\to\infty} (\mathbb{P}(A) + \mathbb{P}(B)) = \lim_{n\to\infty} \mathbb{P}(A) + \lim_{n\to\infty} \mathbb{P}(B) = e^{-\lambda T}\frac{(\lambda T)^k}{k!} + 0 = e^{-\lambda T}\frac{(\lambda T)^k}{k!}$$

where we have made use of (5.6). Consequently, $X \sim \text{Pois}(\lambda T)$ as was asserted.

[a] We remark that, unlike trials where occurrences/outcomes happen in discrete time, we are here working with time on a continuum and so, in principle, occurrences can happen at any time t within the interval $(0, T]$.

Exercise 5.10:

Suppose that people arrive randomly at a post office and, as we discussed in class, it is reasonable to model the number of people that arrive in a given period of time by a Poisson random variable.

1. If the post office opens at 10AM and it is expected that 6 people will arrive between 10AM and noon. What is the probability that at least two people will arrive between 10AM and noon?
2. What is the probability that no one arrives between 10AM and 11AM?
3. What is the expected number of people that will arrive between 10AM and 11AM

5.2.5 Some more exercises

Exercise 5.11:

Let X be a random variable which you may assume to be defined on a countable sample space.

1. If f and g are real-valued functions such that $f(x) \leq g(x)$ for all $x \in \mathbb{R}$, show that
$$\mathbb{E}(f(X)) \leq \mathbb{E}(g(X)).$$
In other words, the expectation is "monotonic".
2. For a real number a, consider the function
$$f_a(x) = \begin{cases} a & \text{if } x \geq a \\ 0 & \text{if } x < a. \end{cases}$$
Show that
$$\mathbb{E}(f_a(X)) = a\mathbb{P}(a \leq X).$$
3. Suppose, additionally, that X is a non-negative random variable, i.e., its range is a subset of $[0, \infty)$. Show that, for any $a > 0$,
$$\mathbb{P}(a \leq X) \leq \frac{\mathbb{E}(X)}{a}$$
by completing the following three steps:
   (a) Argue that
   $$f_a(x) \leq g(x)$$
   where
   $$g(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0. \end{cases}$$

(b) Using the assumption that X is non-negative, show that

$$\mathbb{E}(X) = \mathbb{E}(g(X)).$$

(c) Use the monotonicity property you demonstrated in the first Item to make the final conclusion.

Given the a Poisson random variable X with parameter λ, show that Markov's inequality guarantees that, for any $a \in \mathbb{N}_+$,

$$\sum_{k=a}^{\infty} \frac{\lambda^k}{k!} \leq \frac{\lambda}{a} e^{\lambda}.$$

5.3 Continuous Random Variables

In the previous section, we studied discrete random variables. As we discussed, in the case that the sample space Ω was countable, all random variables are necessarily discrete. We also gave examples of discrete random variables on uncountable subspaces. In this subsection, we discuss a new class of random variables which we will call continuous. Before we are able to do this, it is first helpful to introduce a "continuous" analogue of the probability mass function.

Definition 5.28. *A function $f : \mathbb{R} \to \mathbb{R}$ is called a probability density function if it satisfies the following two properties:*

1. *It is non-negative, i.e., $f(x) \geq 0$ for all $x \in \mathbb{R}$.*

2. *It has unit total mass in the sense that*

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

Remark 5.29. As a technical matter, we should also require that probability density functions be somewhat "nice". For example, you can think of f as continuous or piecewise continuous and these are all that we will see in this course. If you do go on to study probability, you will learn that they can be much less well behaved than piecewise continuous. The correct technical requirement is called "Lebesgue measurable".

Remark 5.30. As you might remember from calculus, integrating from $-\infty$ to ∞ has to be done with some care. We shall interpret these integrals as improper Riemann integrals (maybe an appendix item?). If you go on to study probability beyond this course, you will see that these integrals should be interpreted in the sense of Lebesgue.

Example 5.23: A Simple Density Function

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

for $x \in \mathbb{R}$. The graph of f is illustrated in Figure 5.6.

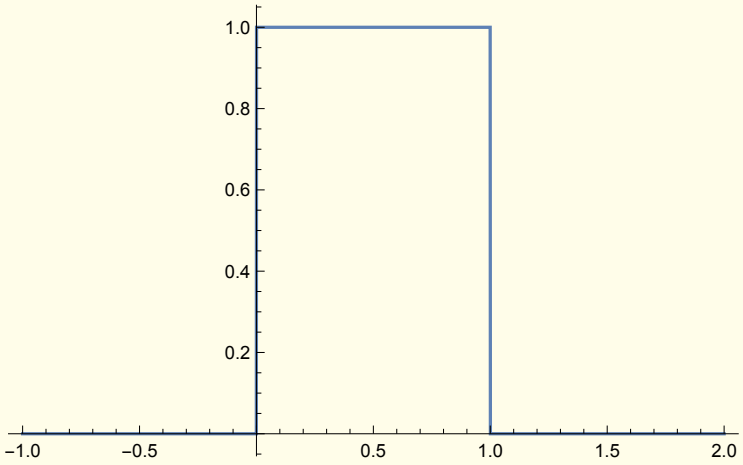


Figure 5.6: The graph of f for $-3\pi < x \leq 3\pi$.

We observe that f takes on only two values: 0 and 1 and so it follows that $f(x) \geq 0$ for all $x \in \mathbb{R}$. Also,

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{1} f(x)\,dx + \int_{1}^{\infty} f(x)\,dx \\ &= \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} 1\,dx + \int_{1}^{\infty} 0\,dx \\ &= 0 + 1 + 0 = 1.\end{aligned}$$

where we have used the fact that f is equal to one on the interval $[0,1]$ and zero on the intervals $(-\infty, 0)$ and $(1, \infty)$. Thus, f is a non-negative function with unit total mass and we may conclude that f is a probability density function.

Example 5.24: Finding C

Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} Ce^{-2x} & 0 \leq x < \infty \\ 0 & x < 0 \end{cases}$$

for $x \in \mathbb{R}$ where C is a constant. We claim that, for an appropriate choice of C, f is a probability density function. To see this, we first observe that f is non-negative provided that $C \geq 0$. Also,

$$\begin{aligned}\int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx \\ &= \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} Ce^{-2x}\,dx \\ &= C\int_{0}^{\infty} e^{-2x}\,dx\end{aligned}$$

To compute this final integral, we investigate its convergence as an improper Riemann integral and, to this end, we observe that

$$\lim_{t\to\infty}\int_{0}^{t} e^{-2x}\,dx = \lim_{t\to\infty} -\frac{e^{-2x}}{2}\Big|_{x=0}^{x=t} = \lim_{t\to\infty}\frac{(-e^{-2t} - (-e^{2(0)}))}{2} = \lim_{t\to\infty}\frac{1(1-e^{-2t})}{2} = \frac{1-0}{2} = \frac{1}{2}.$$

Consequently, the improper Riemann integral $\int_0^\infty e^{-2x}\,dx$ does exist and is equal to $1/2$. For f to be a probability density function, we must have

$$1 = \int_{-\infty}^{\infty} f(x)\,dx = C\int_0^\infty e^{-2x}\,dx = C\lim_{t\to\infty}\int_0^t e^{-2x}\,dx = \frac{C}{2}.$$

and this can be arranged by choosing $C = 2$. Consequently,

$$f(x) = \begin{cases} 2e^{-2x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is a *bona fide* probability density function.

Exercise 5.12: Are these probability density functions?

If possible, determine a constant C for which the following functions are probability density functions. If it is not possible to find such a C, explain why.

1. $$f(x) = \begin{cases} Cx(1-x) & 0 \leq x \leq 1 \\ 0 & \text{else.} \end{cases}$$

2. $$f(x) = Ce^{-|x|}$$

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

4. $$f(x) = \begin{cases} C\sin(x) & 0 \leq x \leq \pi \\ 0 & \text{else.} \end{cases}$$

5. $$f(x) = \begin{cases} C\sin(x) & 0 \leq x \leq 3\pi \\ 0 & \text{else.} \end{cases}$$

Armed with the notion of probability density function, we are able to introduce a new class of random variables.

Definition 5.31. *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. A random variable X on Ω is said to be continuous if there exists a probability density function $f = f_X$ for which*

$$\mathbb{P}(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\,dx$$

for all $x_1 < x_2$. We say that f_X is the probability density function (p.d.f.) of X.

Remark 5.32. You should not think of continuous random variables as being continuous functions, nor should you think of them as having continuous densities – both are usually not true. Though it is still not the reason that the term continuous is used (which comes from measure theory), it is perhaps best to think of continuous random variables as those which take values on the continuum, e.g., their ranges can be full intervals and not just discrete points. Given this possibility of confusion, some authors (including Chung [14]) refrain from using the terminology "continuous random variable" and instead call them "random variables with densities". While I agree that the term

"continuous random variable" is confusing and anachronistic, it is used commonly throughout the literature and I don't want to put you at a disadvantage by not using it.

Continuous random variables are used to model quantities that take values on the continuum, i.e., these are random variables whose ranges can fall anywhere within an interval (or intervals) of real numbers. For example, height, weight, density, energy, pressure, position, direction, distance, and time can all be modeled by continuous random variables. As we did for discrete random variables, we will consider several important types of continuous random variables. The simplest of these is introduced as follows.

Definition 5.33. *Let X be a random variable whose range is the interval $[a, b]$ (with $a < b$). We say that X is uniform on $[a, b]$ provided that it has probability density function*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise.} \end{cases}$$

In this case we write $X \sim Unif([a, b])$.

Uniform random variables are the continuum analogues of the discrete uniform random variables of Example WHICH. Given a uniform random variables X on $[a, b]$ and any subinterval $I = [x_1, x_2] \subseteq [a, b]$, observe that

$$\mathbb{P}(X \in I) = \int_{x_1}^{x_2} \frac{1}{b-a}\, dx = \frac{x_2 - x_1}{b-a} = \frac{\ell(I)}{\ell([a,b])}$$

where $\ell(I)$ denotes the length of the interval I. In fact, it holds that for any reasonable set $I \subseteq [a, b]$,

$$\mathbb{P}(X \in I) = \frac{\ell(I)}{\ell([a,b])}$$

where $\ell(I)$ is a measure of I which straightforwardly generalizes the notion of length. Thus the probability of the event $\{X \in I\}$ is simply the proportion of the length of I to the length of the total range $[a, b]$ of X; it is insensitive to the location of I within $[a, b]$ (*do you see why?*).

In terms of modeling real-life situations, uniform random variables are those that describe numerical outputs of experiments where a continuum of values can be attained and where every interval of those values is assigned the same (i.e., uniform) probability. For example, if we throw a marble within a box of dimension $[a, b] \times [c, d]$ in such a way that the random landing position falls within a region A with probability proportional to the area of A, then the marble's ordinate (x-position) is a uniform random variables $X \sim \text{Unif}([a, b])$. Here is another example.

Example 5.25: Ship Heading

A traveling ship chooses a direction Θ (made with the x-axis) to travel where $\Theta \sim \text{Unif}([-\pi/2, \pi/2])$. With this model, we can compute several probabilities concerning the heading/location of the ship. For example, since

$$f_\Theta(\theta) = \frac{1}{\pi/2 - (-\pi/2)} = \frac{1}{\pi}$$

for $\theta \in [-\pi/2, \pi/2]$, the probability that a heading is chosen at most $\pi/3$ radians from the x axis is

$$\mathbb{P}\left(|\Theta| \leq \frac{\pi}{3}\right) = \mathbb{P}\left(-\frac{\pi}{3} \leq \Theta \leq \frac{\pi}{3}\right) = \int_{-\pi/3}^{\pi/3} \frac{1}{\pi}\, d\theta = \frac{\ell([-\pi/3, \pi/3])}{\pi} = \frac{2}{3}.$$

We can also answer questions concerning the ship's position after traveling in the direction Θ at a constant speed. For instance, if a ship travels at constant speed $v = 10km/hr$ in the direction Θ, the probability that, after 2 hours, the ship has traveled more than $10km$ from the y-axis can be computed as follows. After $t = 2$ hours, the distance the ship has traveled from the y-axis is $X = vt\cos(\Theta) = 20\cos(\Theta)$ measured in

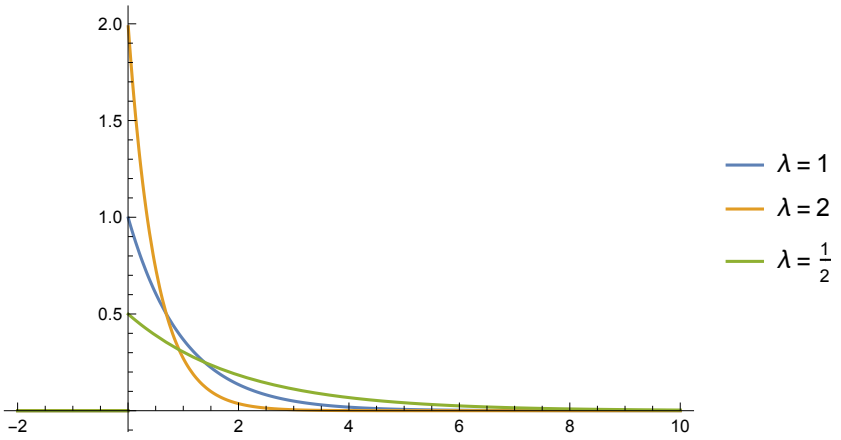


Figure 5.7: Densities of the exponential random variable

km. Thus,

$$\mathbb{P}(X \geq 10) = \mathbb{P}(20\cos(\Theta) \geq 10) = \mathbb{P}(\cos(\Theta) \geq 1/2) = \mathbb{P}\left(-\frac{\pi}{3} \leq \Theta \leq \frac{\pi}{3}\right) = \frac{2}{3};$$

here, we have used the fact that $\cos(\theta) \geq 1/2$ if and only if $-\pi/3 \leq \theta \leq \pi/3$ for $\theta \in [-\pi/2, \pi/2]$. In Subsection 5.4.1, we shall study functions of random variables in some more depth.

Before we introduce our next major example of a continuous random variable, called the exponential random variable, let's take care of a quick mathematical fact.

Lemma 5.34. *Let $\lambda > 0$. Then*

$$f(x) = f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

is a probability density function.

Proof. By inspection, we see that f_λ is non-negative and

$$\int_{-\infty}^{\infty} f_\lambda(x)\,dx = \int_0^{\infty} \lambda e^{-\lambda x}\,dx = \lim_{t\to\infty} \int_0^t \lambda e^{-\lambda x}\,dx = \lim_{t\to\infty} 1 - e^{-\lambda t} = 1$$

□

Definition 5.35. *Let $\lambda > 0$. A continuous random variable X is called an exponential random variable with parameter λ if it has the distribution*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for $x \in \mathbb{R}$. We shall also say that X is exponentially distributed and write $X \sim Exp(\lambda)$.

Given that f_X only assigns positive probabilities to events $\{X \in I\}$ where I contain positive numbers, we see that exponential random variables are non-negative. Figure 5.7 illustrates the density of the exponential random variable for $\lambda = 1, 2$ and $1/2$. As we see, the larger λ is, the more concentrated the density at the low end (near zero). The density becomes more spread out as λ decreases. We shall soon see how this can be interpreted through mean and variance.

Exponential random variables are used to model the waiting time until an event occurs. For example, the lifetime of computers, light bulbs, cars, and appliances are often modeled by exponential random variables.

Example 5.26: Light Bulb Lifetime

A company manufactures light bulbs and keeps good track of their lifetime (the time from the manufacture date until the light goes out). They find that this lifetime can be modeled as a random variable $T \sim \text{Exp}(\lambda)$ where $\lambda = 1/2$ has units of $(\text{year})^{-1}$. With this, we can answer various questions about the probability of the lifetime of a given light bulb. For example, the probability that a light bulb lasts more than three years is

$$\mathbb{P}(T > 3) = \int_3^\infty \frac{1}{2} e^{-x/2}\, dx = \lim_{t\to\infty} -e^{-x/2}\Big|_3^t = e^{-3/2} \approx 0.223.$$

Similarly, the probability that the random variable lasts at most two years is

$$\mathbb{P}(T \leq 2) = 1 - e^{-2/2} = 1 - 1/e \approx 0.632.$$

Exercise 5.13: Time of Arrival

For a bank that opens at 9AM, the arrival time of the first customer can be modeled by $T = 9 + X$ where X (measured in hours) is an exponential random variable of parameter $\lambda = 0.1$.

1. What is the probability that the first customer will arrive after 10AM?
2. What is the probability that the first customer will arrive between 9:00AM and 9:20AM?

Remark 5.36. Warning: Students of probability will often confuse Poisson random variables with exponential random variables. As Poisson random variables are discrete (taking only integer values) and exponential random variables are continuous (taking a values on a continuum), they are different in nature and model different things. Still, there is a connection between them: The Poisson random variable counts the number of rare events in an interval of time and the exponential random variable (with the same parameter) models the amount of time between successive (rare) events.Come back and reference this. In this way, these random variables are "born" from the same model.

Our next random variable is called the normal random variables and it is central to probability and statistics (and all of mathematics, for that matter). As we shall see, its density is a scaled version of e^{-x^2} and does not have a simple antiderivative in closed form. For this reason, before we introduce the normal random variable, we first treat a lemma which will be very helpful to our analysis.

Lemma 5.37.

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}.$$

Proof. Observe that

$$I = \int_{-\infty}^{\infty} e^{-x^2}\, dx = \int_{-\infty}^{\infty} e^{-y^2}\, dy$$

and therefore

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-x^2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2}\, dy\right).$$

Using the linearity of the integral (e.g., that $\int \alpha f = \alpha \int f$) and the Fubini-Tonelli theorem, we have

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-x^2}\, dx \right) e^{-y^2}\, dy \\ &= \int_{-\infty}^{\infty} \left(e^{-y^2} \int_{-\infty}^{\infty} e^{-x^2}\, dx \right) dy \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{-y^2} e^{-x^2}\, dx \right) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2} e^{-y^2}\, dx\, dy \\ &= \int_{\mathbb{R}^2} e^{-(x^2+y^2)}\, dA. \end{aligned}$$

We now make a change to polar coordinates (r, θ) by putting $x = r\cos(\theta)$, $y = r\sin(\theta)$ so that the integration domain in polar coordinates is $\{(r,\theta) : 0 \leq r < \infty,\ 0 \leq \theta \leq 2\pi\}$, and the area element $dA = dx\, dy$ becomes $r\, d\theta\, dr$. Under this change to polar coordinates, we have

$$\begin{aligned} I^2 &= \int_0^{\infty} \int_0^{2\pi} e^{-((r\cos(\theta))^2 + (r\sin(\theta))^2)} r\, d\theta\, dr \\ &= \int_0^{\infty} \int_0^{2\pi} e^{-r^2} r\, d\theta\, dr \\ &= \int_0^{\infty} 2\pi e^{-r^2} r\, dr \\ &= \pi \int_0^{\infty} 2re^{-r^2}\, dr \end{aligned}$$

where we have used the fact that $\cos^2(\theta) + \sin^2(\theta) = 1$. Since $\frac{d}{dr}(-e^{-r^2}) = 2re^{-r^2}$, we have

$$\int_0^{\infty} 2re^{-r^2}\, dr = \lim_{t\to\infty} \int_0^t 2re^{-r^2}\, dr = \lim_{t\to\infty} (-e^{-r^2})\Big|_0^t = \lim_{t\to\infty} (1 - e^{-t^2}) = 1 - 0 = 1.$$

Thus

$$I^2 = \pi \int_0^{\infty} 2re^{-r^2}\, dr = \pi \cdot 1 = \pi$$

or, equivalently, $I = \sqrt{\pi}$. □

With the help of the above lemma, we may now consider an important family of probability density functions. These are called normal densities and are the subject of the following corollary.

Corollary 5.38. *Let μ and σ be real numbers for which $\sigma > 0$. Consider the function*

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

defined for $x \in \mathbb{R}$. Then f is a probability density function.

Proof. Because the function f is non-negative, we only need to show that

$$\int_{-\infty}^{\infty} f(x)\, dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 1.$$

To compute the integral, let's make the change of variables $y = (x-\mu)/\sqrt{2\sigma^2}$ so that $(x-\mu)^2/2\sigma^2 = y^2$, $dx = \sqrt{2\sigma^2}dy$ and the integration of x from $-\infty$ to ∞ is an integration of y over the same domain. Doing this we find that

$$\begin{aligned}\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-y^2}\sqrt{2\sigma^2}\,dy \\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-y^2}\,dy \\ &= \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1.\end{aligned}$$

□

We are now in a position to define the normal random variable.

Definition 5.39. *Let μ and σ be real numbers with $\sigma > 0$. A continuous random variable X is said to be normal with parameters μ and σ^2 if it has probability density function*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for $x \in \mathbb{R}$. We shall also say that X is normally distributed in this case and write $X \sim \mathcal{N}(\mu, \sigma^2)$. In the case that $\mu = 0$ and $\sigma^2 = 1$, we call X standard normal. We shall often denote standard normal random variables by Z.

Figure 5.8 illustrates the normal density in several cases of μ and σ. Observe that these densities are symmetric about $x = \mu$ and seem to spread out as σ increases/concentrate as σ decreases.

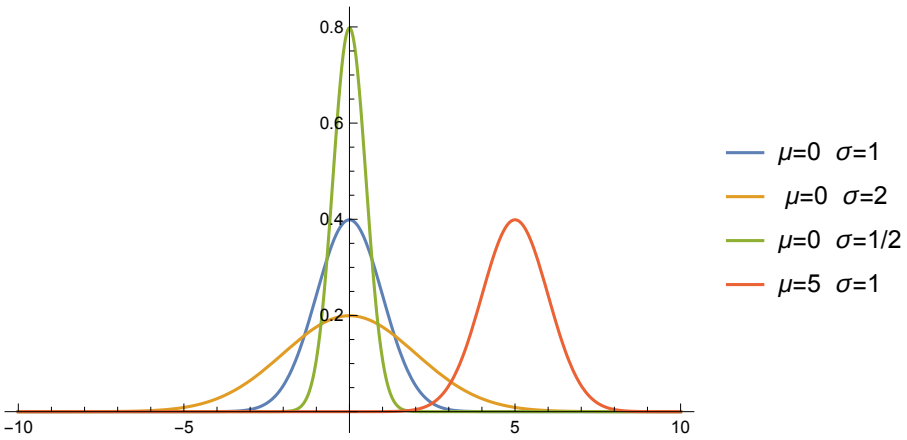


Figure 5.8: Density of Normal Random Variables

The concept of a normal random variable was introduced by Abraham DeMoivre to approximate probabilities associated to the binomial random variable (for large n). Due to a celebrated theorem, called the central limit theorem, we shall see that the normal random variable provides a fantastic model for many natural phenomena. Examples of this are often statistical in nature (those produced by having large numbers/averaging) and include measurements made in physical quantities such as the height of people and the momentum of molecules in gasses.

Example 5.27: Statistical Mechanics

The branch of physics called statistical mechanics is concerned with studying the macroscopic properties of a large number of particles which undergo, often complicated, microscopic interactions. For example, one

can ask about the macroscopic properties (such as pressure, temperature, density) of a gas containing a large number (on the order of moles, $\approx 6.022 \times 10^{23}$) of molecules which collide with each other often. As keeping track of the dynamics of such a large number of particles is practically impossible[a], statistical mechanics aims to model the problem probabilistically by answering equations concerning what a given particle is doing on average and, equivalently[b], the proportion of particles exhibiting a certain behavior.

This theory was originally developed by several important physicists and mathematicians (including Boltzman, Maxwell, Gibbs, Birkhoff, and von Neumann), it is well understood that the location and velocity of particles in a so-called ideal gas can be modeled by normal random variables.

As a toy model, let's assume that we have an ideal gas for which the x-velocity (velocity in the x-direction) of particles can be modeled by a standard normal random variable $V \sim \mathcal{N}(0,1)$ which we take to be measured in centimeters per second. We can ask: What is the fraction of particles whose x-velocity can be found between -1 and 1 cm/s? Using the density of the normal random variable with $\mu = 0$ and $\sigma = 1$, we have

$$\mathbb{P}(-1 \leq V \leq 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx.$$

As we already saw, computing these types of so-called Gaussian integrals is not straightforward. A common way is to introduce the function

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

which is equivalently the cumulative distribution function the standard normal random variable. Though the values of this function cannot be computed exactly by hand, they can be approximated to any desired accuracy. In days past (before modern computes), you would be handed a table of the values which you would then use to approximate the answer; today, of course, we use a computer. With the aid of Mathematica, we compute

$$\begin{aligned}\mathbb{P}(-1 \leq V \leq 1) &= \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx - \int_{-\infty}^{-1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\ &= \Phi(1) - \Phi(-1) \approx 0.841345 - 0.158655 = 0.682689.\end{aligned}$$

Thus, for this ideal gas, approximately 68 percent of particles, at any given time, can be found to have x-velocity between -1 and 1 cm/s.

Of course, we could instead ask about the fraction of particles with x-velocity in any interval, not just that between -1 and 1. Curiously, we could ask about the fraction of particles whose x-velocity is more than the speed of light, $c \approx 3.0 \times 10^{10} cm/s$. This is

$$\mathbb{P}(V > c) = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx$$

and is necessarily a strictly positive number because $e^{-x^2/2}/\sqrt{2\pi} > 0$ on the interval $[c, \infty)$. Perhaps this should worry us because Einstein's theory of relativity says that nothing moves faster than the speed of light. In fact, using the same ideas as we used to prove Lemma 5.37, you can show that this integral is bounded above by $e^{-c^2/4}$ which is roughly 10^{-200}, a number smaller than the machine precision of every computer. In fact, you could select with much greater probability a single particle by searching through all particles in the observable universe [2]. I think you might agree that this probability, though positive, is effectively zero. Appropriately so, Mathematica computes $\mathbb{P}(V > c) = 1 - \Phi(c) = 0$.

[a] It would involve solving a non-linear coupled system of $\approx 36.132 \times 10^{23} = 6$ moles of differential equations

[b] This equivalence is, in fact, a fundamental postulate of the theory called the ergodic hypothesis. The article [7] gives a fascinating account of the history of the ergodic hypothesis and related mathematical theorems of von Neumann and Birkhoff.

For computations involving normal random variables which are not standard normal, the following proposition is

very helpful.

Proposition 5.40. *Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Then, $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if*

$$Z := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

In other words, every normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ is of the form $X = \sigma Z + \mu$ where Z is standard normal.

We shall not spend the time to labor through the proof of this proposition now because, as we will see, it becomes an easy exercise once we start talking about functions of continuous random variables in terms of cumulative distribution functions. For now, let's use the proposition to estimate the height of adults in the following example.

Example 5.28: Modeling Human Height

Human height can be approximated/modeled by a normal random variable H (measured in centimeters) with $\mu = 175$ and $\sigma^2 = (7.5)^2$. For example, we can ask what fraction of adult humans have height less than 175 cm (approximately 5′8”)? This is,

$$\mathbb{P}(H \leq 175) = \int_{-\infty}^{175} \frac{1}{\sqrt{2\pi}(7.5)} e^{-\frac{(x-175)^2}{2(7.5)^2}}\, dx$$

an integral that we might not want to compute (though you can through a change of variables). If instead, we appeal to the proposition above, we see that $H = (7.5)Z + 175$ where Z is standard normal, we see that

$$\mathbb{P}(H \leq 175) = \mathbb{P}((7.5)Z + 175 \leq 175) = \mathbb{P}((7.5)Z \leq 0) = \mathbb{P}(Z \leq 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \frac{1}{2}$$

where we have used the fact that $f_Z(x) = e^{-x^2/2}/\sqrt{2\pi}$ is an even function (i.e., $f_Z(x) = f_Z(-x)$) so that

$$1 = \int_{-\infty}^{\infty} f_Z(x)\, dx = 2\int_{0}^{\infty} f_Z(x)\, dx.$$

Thus, in this model, the fraction of adult humans below 175 centimeters is 1/2. We can also compute, for instance the fraction of adult humans with height above $200cm$ ($\approx 6'7$”). We find that

$$\mathbb{P}(H > 200) = \mathbb{P}((7.5)Z + 175 > 200) = P(Z > 10/3) = 1 - \mathbb{P}(Z \leq 10/3) = 1 - \Phi(10/3) \approx 0.00042906.$$

Exercise 5.14: Normal Copper Tubes

A company produces copper tubes. Through much statistical analysis, they are able to model the height H and diameter D of these copper tubes as normal random variables. Specifically, $H \sim \mathcal{N}(100, 0.4)$ and $D \sim \mathcal{N}(1, 0.001)$ with units measured in centimeters. Using software:

1. Plot the probability density functions for H and D.
2. Approximate, to three decimal places, $\mathbb{P}(H > 99)$.
3. Approximate, to three decimal places, $\mathbb{P}(D > 0.9)$.
4. Assuming that the events $\{H \in I\}$ and $\{D \in J\}$ are independent for any intervals $I, J \subseteq \mathbb{R}$, compute the probability of the event that a given tube is more than 99 centimeters long and less than 0.9 centimeters in diameter.

5.3.1 Expectation of Continuous Random Variables

In looking at Figure 5.8, you probably suspect, that the parameters μ and σ^2 for a normal random variable coincides with the random variable's mean and variance, respectively. Your suspicion is correct. To make this interpretation, however, we first need a reasonable way to understand expectation for continuous random variables. Given a continuous random variable X on a (necessarily uncountable sample space) Ω, let's formally consider the sum

$$\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)$$

used in Definition 5.4. In contrast to the case in which Ω is countable (so that X is necessarily discrete), the above sum is problematic because, while the probability of singleton events are now zero, the sum would be taken over the uncountable set Ω which we do not know how to do. If you take probability beyond this course (or a course in measure theory), you will learn that this sum is adroitly replaced by a type of integral, called the Lebesgue integral. While we lack the machinery here to write down the (correct) analogue of Definition 5.4, we are able to understand the expectation of continuous random variables in terms of their probability density functions analogously to the way that we have understood expectation for discrete random variables in terms of their mass functions. As we have already seen, going from the discrete setting to the continuous, mass functions are repaced by density functions and sums are replaced by integrals. Thus, we expect that the formula

$$\mathbb{E}(X) = \sum k \cdot p_X(k)$$

of Theorem 5.14 when X is discrete to be replaced by

$$\mathbb{E}(X) = \int k f_X(k)\,dk \qquad \text{or equivalently} \qquad \mathbb{E}(X) = \int x f_X(x)\,dx$$

when X is continuous. As the following theorem (which you can take as a definition in the continuous setting) shows, this is indeed the case.

Theorem 5.41. *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. To each[3] random variable X, we can associate a number $\mathbb{E}(X)$ called the expectation of X (or expected value or mean). This association is positivity-preserving in the sense that $\mathbb{E}(X) \geq 0$ whenever X is non-negative. Further, it is linear in the sense that $\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$ for random variables X and Y and constants α and β. In the case that Ω is countable, $\mathbb{E}$ is that given by Definition 5.4. In the case that X is discrete, $\mathbb{E}(X)$ is characterized by Theorem 5.14. Finally, in the case that X is continuous with probability density function f_X, we have the following characterization:*

1. *If $x \mapsto x f_X(x)$ is non-negative or*

$$\int_{-\infty}^{\infty} |x|\, f_X(x)\,dx < \infty,$$

then $\mathbb{E}(X)$ is defined (but is possibly infinite) and

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx. \tag{5.10}$$

2. *More generally, if $\varphi : \mathbb{R} \to \mathbb{R}$ is a reasonable[4] function for which $x \mapsto \varphi(x) f_X(x)$ is non-negative or*

$$\int_{-\infty}^{\infty} |\varphi(x)|\, f_X(x)\,dx < \infty,$$

then

$$\mathbb{E}(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x) f_X(x)\,dx.$$

[3] Technically, for $\mathbb{E}(X)$ to be defined we require that $X : \Omega \to \mathbb{R}$ is non-negative (in which case $\mathbb{E}(X) = \infty$ is possible) or that an "absolute summability" condition akin to that in Definition 5.4 holds. In this course, we shall only study random variables that satisfy one of these conditions and so you don't need to worry about such technical matters here.

[4] We must require that φ be Lebesgue measurable. This is a technical condition that you need not worry about in this course. Essentially, every function that can be written down is Lebesgue measurable.

Remark 5.42. As in the discrete setting, the set of real numbers x for which $f_X(x) > 0$ coincides essentially with the range of X, $R(X)$. For this reason, we can write

$$\mathbb{E}(\varphi(X)) = \int_{R(X)} \varphi(x) f_X(x)\, dx$$

which is analogous to (5.4) of Theorem 5.14 for discrete random variables.

Remark 5.43. Since $f_X(x)$ is non-negative, the above condition that $x \mapsto x f_X(x)$ is non-negative is equivalently the condition that $f_X(x) = 0$ whenever $x < 0$. Do you see why? Along the lines of the previous remark, this is precisely the case when the random variable X is itself non-negative. In this case,

$$\mathbb{E}(X) = \int_0^\infty x f_X(x)\, dx$$

which can be infinite in our interpretation. More generally, the condition that $x \mapsto \varphi(x) f_X(x)$ is non-negative can be met in various ways. Notice that it is automatically satisfied when $\varphi(x)$ is itself non-negative. In particular,

$$\mathbb{E}(X^2) = \int_{-\infty}^\infty x^2 f_X(x)\, dx$$

is always defined though it can be infinite. It is a worthwhile exercise to find density functions f_X for non-negative random variables (so that $f_X(x) = 0$ when $x < 0$) that have infinite means or have finite means and infinite second moment/infinite variance. Such density functions (and their random variables) are said to have heavy tails.

Let's now use Theorem 5.41 to discuss variance and moments. For a random variable X with finite mean $\mu = \mathbb{E}(X)$, we recall that

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

In the broader interpretation of $\mathbb{E}$ described in Theorem 5.41, this identity (and definition of variance) always holds though it can be infinite. As in the discrete setting, a random variable has finite variance if and only if $\mathbb{E}(X^2)$ is finite. In general, we say that a random variable X has a finite moment of order n provided that $\mathbb{E}(|X|^n) < \infty$ and in this case $\mathbb{E}(X^n)$ makes sense and is called the n-th moment of X. In the case that X is continuous with density function f_X, the variance of X is given by

$$\text{Var}(X) = \mathbb{E}((X-\mu)^2) = \int_{-\infty}^\infty (x-\mu)^2 f_X(x)\, dx$$

where

$$\mu = \mathbb{E}(X) = \int_{-\infty}^\infty x f_X(x)\, dx.$$

For $n \in \mathbb{N}$, we have

$$\mathbb{E}(X^n) = \int_{-\infty}^\infty x^n f_X(x)\, dx$$

for the nth moment of X. Further, the variance can be computed using the identity

$$\text{Var}(X) = \mathbb{E}(X^2) - \mu^2 = \int_{-\infty}^\infty x^2 f_X(x)\, dx - \mu^2 = \int_{-\infty}^\infty x^2 f_X(x)\, dx - \left(\int_{-\infty}^\infty x f_X(x)\, dx\right)^2$$

Example 5.29: Uniform Mean and Variance

In this example, we compute the mean and variance a uniform random variable $X \sim \text{Unif}([a,b])$ for $a < b$.

For this random variable, we recall that its distribution function is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}.$$

Therefore,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a}\frac{x^2}{2}\Big|_a^b = \frac{1}{b-a}\frac{b^2-a^2}{2} = \frac{b+a}{2}$$

where we have used the fact that the difference of squares can be factored as $b^2 - a^2 = (b-a)(b+a)$. Interpreting the mean $\mathbb{E}(X)$ as the center of mass of f_X, this makes perfect sense because $(b+a)/2$ is precisely the midpoint between a and b. To compute the variance, let's first compute the second moment of X. We have

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_a^b \frac{x^2}{b-a}\,dx = \frac{1}{b-a}\frac{b^3-a^3}{3} = \frac{b^2+ab+a^2}{3}$$

where we have used the factorization of two cubes, $b^3 - a^3 = (b-a)(b^2+ab+a^2)$. Consequently,

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{4(b^2+ab+a^2)}{12} - \frac{3(b^2+2ab+b^2)}{12} \\ &= \frac{b^2-2ab+a^2}{12} \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

As the variance measure the spread of f_X away from the mean, we see that the above formula makes good intuitive sense. When b is close to a, the mass of f_X is concentrated on the small interval $[a,b]$ which looks like a spike and, in this case, $\text{Var}(X) = (b-a)^2/12$ is small. If b is far from a, this density is spread out and quite appropriately, the variance is large in this case. Figure 5.9 illustrates this observation.

Figure 5.9: Examples of the uniform density when $|a-b| < 1$ and $|a-b| > 1$.

.

Example 5.30: Exponential Mean and Variance

Consider an exponential random variable X with parameter $\lambda > 0$. As we illustrated in Figure 5.7, the density with $\lambda = 2 > 1$ has its mass more concentrated near $x = 0$ and the density with $\lambda = 1/2 < 1$ spreads is mass out further away from $x = 0$. In view of these observations and by using our geometric intuition for the mean as the center of mass of the density, we could conjecture that the mean is inversely proportional to the parameter λ. Let's compute the mean to test our conjecture. We have

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\,dx = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx = \lim_{t\to\infty}\int_0^t x\lambda e^{-\lambda x}\,dx$$

In looking at the integrands $x\lambda e^{-\lambda x}$ I don't immediately see an antiderivative with which I could invoke the fundamental theorem of calculus to do the computation in one fell swoop. This does however seem like a good candidate for integration by parts. Recall the integration by parts formula

$$\int_a^b u\,dv = uv\Big|_a^b - \int_a^b v\,du.$$

If we set $u = x$, $dv = \lambda e^{-\lambda x}$, we find that $du = 1$ and $v = -e^{-\lambda x}$, we find

$$\begin{aligned}\int_0^t x\lambda e^{-\lambda x}\,dx &= x(-e^{-\lambda x})\Big|_0^t - \int_0^t (-e^{-\lambda x})\,dx\\ &= -te^{-\lambda t} - 0 + \int_0^t e^{-\lambda x}\,dx\\ &= -te^{-\lambda t} + \left(-\frac{1}{\lambda}e^{-\lambda x}\right)\Big|_0^t\\ &= \frac{1}{\lambda} - \left(\frac{1}{\lambda} + t\right)e^{-\lambda t}\end{aligned}$$

for each $t > 0$. Upon noting that

$$\lim_{t\to\infty}\left(\frac{1}{\lambda} + t\right)e^{-\lambda t} = 0$$

which can be seen using L'Hôpital's rule, we find that

$$\mathbb{E}(X) = \lim_{t\to\infty}\int_0^t xe^{-\lambda x}\,dx = \lim_{t\to\infty}\left(\frac{1}{\lambda} - \left(\frac{1}{\lambda} + t\right)e^{-\lambda t}\right) = \frac{1}{\lambda} - \lim_{t\to\infty}\left(\frac{1}{\lambda} + t\right)e^{-\lambda t} = \frac{1}{\lambda}.$$

which is exactly what we expected: the mean is inversely proportional to λ. By a similar computation (one you should do for yourself – it used integration by parts twice), we find

$$\mathbb{E}(X^2) = \frac{2}{\lambda^2}$$

and so

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Exercise 5.15: The Gamma Function

Throughout the course, we have made essential use of the factorial $n!$ which we've defined for natural numbers $n \in \mathbb{N} = \{0, 1, \dots\}$. In this exercise, we are going to study Euler's so-called Gamma function which gives us

a useful way to extend $n!$ to values of n which are not integers. To this end, for $\alpha > 0$, define

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\, dy.$$

We should, perhaps, worry a little about the convergence of this improper Riemann integral. You should feel free to work out the details if you want, however, you can take for granted that this improper Riemann integral converges whenever $\alpha > 0$. Do the following:

1. Compute $\Gamma(1)$ by evaluating the integral.
2. Show that, for $\alpha > 0$, $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. Hint: Use integration by parts and the fact that $\frac{d}{dy} y^\alpha = \alpha y^{\alpha-1}$.
3. Use the preceding two items to show that $\Gamma(n+1) = n!$ for every $n \in \mathbb{N}$.

Exercise 5.16: The Gamma Distribution

Let $\alpha > 0$ and $\lambda > 0$. A continuous random variable X is said to have the Gamma distribution with associated parameters α and λ if its density function is given by

$$f_X(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for $x \in \mathbb{R}$. It is reasonable to call X a Gamma random variable with parameters λ and α. As you will show in your next homework, when $\alpha = n$ is an integer, this random variable models the amount of time one has to wait until n independent and rare events occur. Of course, an exponential random variable models the waiting time until a single event occurs and this should make sense if you observe that the above density function is exactly that of the exponential when $\alpha = n = 1$.
Let X have the Gamma distribution with $\alpha, \lambda > 0$ and let $f_X(x)$, defined above, be its associated density function.

1. Verify that f_X is indeed a probability density function, i.e., verify that it is non-negative and its integral from $-\infty$ to ∞ is 1.
2. Use software to plot $f_X(x)$ when $(\alpha, \lambda) = (1, 1)$, $(\alpha, \lambda) = (4, 1)$, $(\alpha, \lambda) = (4, 1/2)$ and $(\alpha, \lambda) = (4, 2)$.
3. Use what you did in the previous Problem to show that
$$\mathbb{E}(X) = \frac{\alpha}{\lambda}$$
4. $$\text{Var}(X) = \frac{\alpha}{\lambda^2}$$
5. Given our interpretation of the mean as the "center of mass" and the variance as a measure of the "spread" of the density function, do these expressions for the mean and variance seem to be consistent with your graphs of density functions?

Next, we compute the mean and variance of normal random variables.

Example 5.31: Normal Mean and Variance

In this example, we compute the mean and variance of a normal random variable X with parameters μ and σ^2. Doing this directly is a somewhat tedious task and we will find it easier to first do the computations for a standard normal random variable $Z \sim \mathcal{N}(0,1)$ and then employ Proposition 5.40. We recall that

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for $x \in \mathbb{R}$. Consequently,

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} x \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\, dx = \lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} x e^{-x^2/2}\, dx.$$

Though this integral can be done directly (using integration by parts), it is a little easier to use a little geometric reasoning.

From single-variable calculus recall that a function $g : \mathbb{R} \to \mathbb{R}$ is said to be *odd* if $g(-x) = -g(x)$ for all x. In this case, the graph of the function has an important symmetry in that, if it is rotated by π radians about the origin, the result is identical to the original graph (this is the same as a flip over the x-axis and then a flip ver the y-axis). In this case, any positive area under the graph is matched with (signed) negative area above the graph and so it follows that

$$\int_{-a}^{a} g(x)\, dx = 0$$

for all $a > 0$.With this fact in mind, it is easy to see that the function $x \mapsto x e^{-x^2/2}$ is an odd function and so

$$\mathbb{E}(Z) = \lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} x e^{-x^2/2}\, dx = \lim_{t\to\infty} 0 = 0.$$

We now move to compute that variance of the standard normal random variable. Upon noting that $\mathbb{E}(Z) = 0$,

$$\mathrm{Var}(Z) = \mathbb{E}(Z^2) = \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} x^2 e^{-x^2/2}\, dx.$$

As we did with the exponential random variable, we shall integrate by parts to perform this computation. To this end, it is helpful to write

$$\int_{-t}^{t} x^2 e^{-x^2/2}\, dx = \int_{-t}^{t} x(x e^{-x^2/2})\, dx$$

because we can easily find an antiderivative of $x e^{-x^2/2}$ and, though parts integration, get rid of the factor x. Writing $u = x$ and $dv = x e^{-x^2/2}$, we find that $v = -e^{-x^2/2}$ (check it) and $du = 1$ so that

$$\begin{aligned}
\int_{-t}^{t} x^2 e^{-x^2/2} &= \int_{-t}^{t} x(x e^{-x^2/2})\, dx \\
&= x(-e^{-x^2/2})\Big|_{-t}^{t} - \int_{-t}^{t} (-e^{-x^2/2})\, dx \\
&= -te^{-t^2/2} - (-t)(-e^{-(-t)^2/2}) + \int_{-t}^{t} e^{-x^2/2}\, dx \\
&= \int_{-t}^{t} e^{-x^2/2}\, dx - 2te^{-t^2/2}
\end{aligned}$$

Consequently,

$$\mathrm{Var}(Z) = \lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} x^2 e^{-x^2/2}\,dx = \lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-x^2/2}\,dx - \lim_{t\to\infty} \frac{2}{\sqrt{2\pi}} t e^{-t^2/2}$$

Though, at first glance, we might suspect that we have just gone in a circle with this calculation, a moment's thought gives us the perspective that the limits of integrals on the right is simply the improper Riemann integral of the density f_Z itself and is one, i.e.,

$$\lim_{t\to\infty} \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-x^2/2}\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1.$$

Upon noting that

$$\lim_{t\to\infty} t e^{-t^2/2} = 0$$

which can be seen as an application of L'Hôpital's rule, we conclude that

$$\mathrm{Var}(Z) = 1 + 0 = 1.$$

So, we have shown that the standard normal random variable Z with parameters 0 and 1 has mean 0 and variance 1. In fact, this is no coincidence.

Proposition 5.44. *Let X be a normal random variable with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then*

$$\mathbb{E}(X) = \mu \qquad \text{and} \qquad \mathrm{Var}(X) = \sigma^2.$$

Thus, we can rightly call the parameters μ and σ^2 with which we introduced the normal random variable as its mean and variance, respectively.

Proof. In view of Proposition 5.40, $X = \sigma Z + \mu$ and so, by the linearity of expectation given in Theorem 5.41, we have

$$\mathbb{E}(X) = \mathbb{E}(\sigma Z + \mu) = \sigma\mathbb{E}(Z) + \mathbb{E}(\mu) = \sigma \cdot 0 + \mu = \mu$$

where we have used our computation above that $\mathbb{E}(Z) = 0$ and the result of Exercise 5.1 (which holds in the continuous random variable setting as well) that the expectation of the non-random random variable μ is μ itself. Using our established fact that $\mathbb{E}(X) = \mu$, we have

$$\mathrm{Var}(X) = \mathbb{E}((X-\mu)^2) = \mathbb{E}((\sigma Z)^2)\mathbb{E}(\sigma^2 Z^2) = \sigma^2\mathbb{E}(Z^2) = \sigma^2 \cdot 1 = \sigma^2$$

where we have again used the linearity of expectation and the fact that the second moment of $Z \sim \mathcal{N}(0,1)$ is 1. □

Exercise 5.17: From one moment to the next

Let Z be a standard normal random variable.

1. If $\varphi : \mathbb{R} \to \mathbb{R}$ is differentiable everywhere[a]
$$\mathbb{E}(\varphi'(Z)) = \mathbb{E}(Z\varphi(Z)).$$
2. Use what you showed above to conclude that
$$n\mathbb{E}(Z^{n-1}) = \mathbb{E}(Z^{n+1})$$
and, in view of the fact that $\mathbb{E}(Z) = 0$ and $\mathbb{E}(Z^2) = 1$, use this to find an expression for $\mathbb{E}(Z^k)$ for an arbitrary integer $k \in \mathbb{N}$.

[a]Note: Technically, assuming the φ is differentiable isn't quite enough because we also want the integrals of $\varphi'(x)e^{-x^2/2}$ and $x\varphi(x)e^{-x^2}/2$ to converge. Assuming that φ and $\varphi'(x)$ are bounded is more than enough.

5.4 Cumulative Distribution Functions

We recall that, for a random variable X on a sample space Ω with probability measure $\mathbb{P}$, the cumulative distribution function associated to X is given by

$$F_X(x) = \mathbb{P}(X \leq x)$$

As we learned at the beginning of this chapter (need a pointer), F_X was necessarily a non-decreasing function having $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. In looking back through the examples at the beginning of this chapter (pointers), we saw several examples for discrete random variables where the graphs of their cumulative distribution functions consisted purely of jump discontinuities. In fact, using Theorem 5.14 we find that

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{\substack{k\leq x \\ k\in R(X)}} p_X(x)$$

for $x \in \mathbb{R}$ whenever X is a discrete random variable with probability mass function p_X. In this form, it is apparent that F_X, for a discrete random variable has jump continuities precisely at the points $x = k \in R(X)$ (with $p_X(k) > 0$)). Away from these points, F_X is constant. Putting our calculus hats on, F_X, for a discrete random variable X is not a particularly nice function – it is discontinuous at every point $x \in R(X)$. By contrast, the CDF's of continuous random variables are much nicer.

Given a continuous random variable X with probability density function f_X we have

$$F_X(x) = \mathbb{P}(-\infty < X \leq x) = \int_{-\infty}^{x} f_X(u)\,du;$$

here, we're integrating over the dummy variable u instead of x so as not to cause confusion. Let's consider this function in three cases of continuous random variables that we know well.

Example 5.32: The Uniform CDF

Let $X \sim \text{Unif}([a,b])$ with density $f_X(x) = 1/(b-a)$ for $a \leq x \leq b$ and $f_X(x) = 0$ otherwise. If $x < a$, we see that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{-\infty}^{x} 0\,du = 0$$

because f_X is zero on the interval $(-\infty, a)$. For $a \leq x \leq b$, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{a}^{x} \frac{1}{b-a}\,du = \frac{x-a}{b-a}$$

and, for $x > b$, we have

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du = \int_{a}^{b} \frac{1}{b-a} = 1.$$

Thus

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

which is illustrated in Figure 5.10.

Let's now take a look at the cumulative distribution function of an exponential random variable.

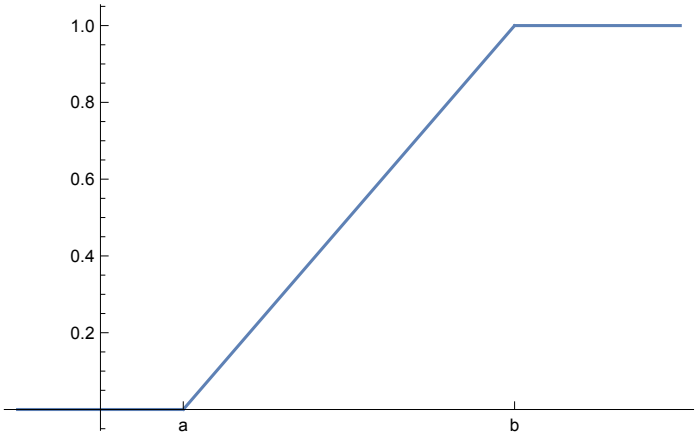


Figure 5.10: CDF of Uniform Random Variable

Example 5.33: The Exponential CDF

Let $X \sim \text{Exp}(\lambda)$ for $\lambda > 0$. Here,

$$f_X(u) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

for $u \in \mathbb{R}$. For $x < 0$,

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = 0$$

since f_X is zero on the interval $(-\infty, x]$. For $x \geq 0$, we then have

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du = \int_{0}^{x} \lambda e^{-\lambda u}\, du = -e^{-\lambda u}\Big|_0^x = 1 - e^{-\lambda x}.$$

Thus

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

for $x \in \mathbb{R}$; we illustrate this in Figure **??**.

Finally, we consider the cumulative distribution function for the normal random variable.

Example 5.34: The Normal CDF

Given $X \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(u-\mu)^2}{2\sigma^2}}\, du$$

for $x \in \mathbb{R}$. In particular, for a standard normal random variable $Z \sim \mathcal{N}(0, 1)$,

$$F_Z(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du = \Phi(z)$$

for $z \in \mathbb{R}$. As we discussed previously, there isn't a more useful simplification of $\Phi(z) = F_Z(x)$ (nor is there for $F_X(x)$) but we can easily make use of computational software to understand its values. This is exactly what I've done in illustrating $F_Z(z) = \Phi(z)$ in Figure 5.11.

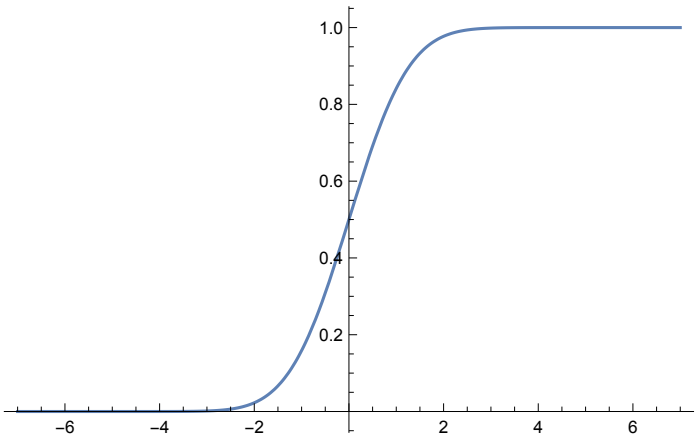


Figure 5.11: The cumulative distribution function $F_Z = \Phi$

In looking at the preceding examples of the three major continuous random variables we've studied, in contrast to discrete random variables, each cumulative distribution function is continuous. Not only are they continuous, but they appear to also be differentiable (having a well-defined slope) at most points. As the following theorem shows, in particular, these observations are representative of continuous random variables in general. The theorem also gives a useful relationship between the probability density function and cumulative distribution function for any continuous random variable.

Theorem 5.45. *Let X be a random variable on a sample space Ω equipped with probability measure $\mathbb{P}$ and let $F_X(x) = \mathbb{P}(X \leq x)$ be the cumulative distribution function associated to X.*

1. *If X is a continuous random variable with density f_X, then F_X can be expressed by*

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\,du,$$

and it is everywhere absolutely continuous. Further, it is differentiable almost everywhere, and satisfies

$$F_X'(x) = f_X(x)$$

at every point at which f_X is itself continuous.

2. *Conversely, if F_X is absolutely continuous and (necessarily) differentiable almost everywhere, then X is a continuous random variable with density*

$$f_X(x) = F_X'(x)$$

defined at each point x at which F_X is differentiable.

Remark 5.46. Though the above theorems are stated correctly, we have not defined the terms "absolutely continuous" or "almost everywhere". Precise definitions of these terms can be found in [8], for example, but you shouldn't worry about them too much. Absolute continuity can be thought of as slightly better than continuity and all examples we will consider, that are continuous, are also absolutely continuous. The term "almost everywhere" technically means "except on a set of measure zero". For reference, finite and countable sets have measure zero. There are, however, some very interesting uncountable sets which also have measure zero. See, for example, the Cantor set.

Though we won't prove this theorem, it should be noted that the statement follows easily from the fundamental theorem of calculus when f_X is piecewise continuous. For more exotic examples of f_X, the theorem follows from the Lebesgue differentiation theorem (see, e.g., [8, Theorem 3.11]). The utility of this theorem is that it allows us to characterize (and move between) continuous random variables using their cumulative distribution functions. We shall employ this result throughout the next section. To whet your appetite, please do the following exercise.

Exercise 5.18: Normal to Standard Normal

In this exercise, you use the above theorem to establish Proposition 5.40. This is the result that allows us to express every normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ in the form $X = \sigma Z + \mu$ where Z is standard normal. To this end, do the following:

1. For $X \sim \mathcal{N}(\mu, \sigma^2)$, set
$$Z' = \frac{X - \mu}{\sigma}$$
and, using the definition of $F_{Z'}(z) = \mathbb{P}(Z' \leq z)$, show that
$$F_{Z'}(z) = \Phi(z)$$
for each $z \in \mathbb{R}$. Hint: First unravel $\{Z' \leq z\}$ and then make a change of variables in the integral involving f_X to simplify things.
2. Differentiating both sides, use the above theorem to conclude that the density of Z' is precisely the standard normal density. Conclude that $Z := Z' \sim \mathcal{N}(0, 1)$ which gives the proposition.

5.4.1 Understanding Functions of Random Variables

In this subsection, we focus our attention on random variables that can be written as functions of other random variables. Though one can develop a very good theory for general random variables, we shall focus our attention on continuous functions of continuous random variables; this focus turns out to have great applicability. As we will see, the best way to get an understanding of a new random variable is to start by looking at its cumulative distribution function. To get a feeling for things, let's first take a look at an example.

Example 5.35: Pizzas!

Suppose that I make pizzas all night whose radii can be modeled by a uniformly distributed continuous random variable $R \sim \text{Unif}([0, 1])$ measured in meters. Note, $R = 1$ meters is a big pizza!. What is the distribution of the area $A = \pi R^2$ of my pizzas? Since $0 \leq R \leq 1$, it appears that $0 \leq A \leq \pi$ and thus, we expect that the random variable A be continuously distribued on the interval $[0, \pi]$. Let's compute the distrubition function of A. For $a < 0$, we have

$$F_A(a) = \mathbb{P}(A \leq a) = 0.$$

For $0 \leq a \leq \pi$,

$$F_A(a) = \mathbb{P}(A \leq a) = \mathbb{P}(\pi R^2 \leq a) = \mathbb{P}(R \leq \sqrt{a/\pi}) = \int_0^{\sqrt{a/\pi}} dr = \sqrt{\frac{a}{\pi}}.$$

And, finally, for $\pi < a$, $F_A(a) = 1$. Putting this all together, we have

$$F_A(a) = \begin{cases} 0 & a < 0 \\ \sqrt{a/\pi} & 0 \leq a < \pi \\ 1 & a \geq \pi \end{cases}$$

for $a \in \mathbb{R}$.Thus, in view of Theorem 5.45,

$$f_A(a) = \frac{d}{da} F_A(a) = \frac{1}{2\sqrt{\pi a}}$$

for $0 < a < \pi$ and 0 otherwise. With this, we can easily compute the mean of A by

$$\mathbb{E}(A) = \int_{-\infty}^{\infty} a f_A(a)\, da = \int_0^{\pi} \frac{a}{2\sqrt{\pi a}}\, dx = \int_0^{\pi} \frac{\sqrt{a}}{2\sqrt{\pi}}\, da = \frac{a^{3/2}}{3\sqrt{\pi}}\Big|_0^{\pi} = \frac{\pi}{3}.$$

This is, of course, no surprise as Theorem 5.41 gives

$$\mathbb{E}(R^2) = \int_{-\infty}^{\infty} r^2 f_R(r)\, dr = \int_0^1 r^2\, dr = \frac{1}{3}$$

and therefore

$$\mathbb{E}(A) = \mathbb{E}(\pi R^2) = \pi \mathbb{E}(R^2) = \frac{\pi}{3}.$$

Example 5.36: Laser Beam

A laser placed at the origin in $\mathbb{R}^2$ shines at a random angle Θ which is assumed to be uniform on the interval $[-\theta_0, \theta_0]$ where $0 < \theta_0 < \pi/2$. A sensor is placed along the line $x = 1$ which, depending on the angle Θ, intercepts the laser at a height Y. This situation is illustrated in Figure 5.12.

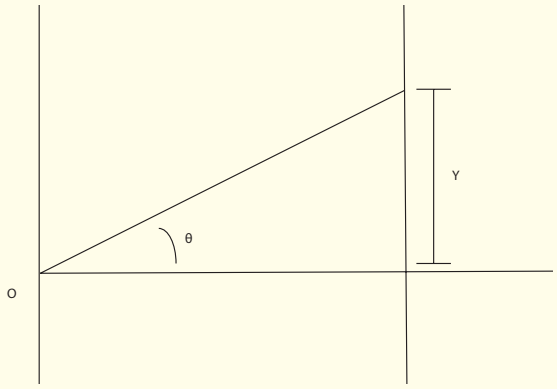


Figure 5.12: Laser light falling incident on wall $x = 1$

We would like to understand the distribution of light along the sensor. In looking at the geometry of Figure 5.12, it appears that Y should not be uniformly distributed even when Θ is uniformly distributed because a range of angles close to zero will occupy less of the line $x = 1$ than will the same range away from 0. Doing a little geometry, we find that

$$Y = \tan(\Theta)$$

where, given that Θ can take values in the interval $[-\theta_0, \theta_0]$, we see that Y can take values on the interval $[-y_0, y_0]$ where $y_0 = \tan(\theta_0)$. Using Theorem 5.41, we can easily compute the mean and variance of Y (without knowing much about Y or how it is distributed) as follows. Under the assumption that $\Theta \sim \text{Unif}([\theta_0, \theta_0])$, we have

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\theta_0} & -\theta_0 < \theta < \theta_0 \\ 0 & \text{otherwise} \end{cases},$$

and therefore

$$\mathbb{E}(Y) = \mathbb{E}(\tan(\Theta)) = \int_{-\theta_0}^{\theta_0} \frac{\tan(\theta)}{2\theta_0}\, d\theta = 0;$$

this can be seen using the fundamental theorem of calculus or the fact that $\tan(\theta)$ is an odd function. Now, given that $\mathbb{E}(Y) = 0$,

$$\text{Var}(Y) = \mathbb{E}(Y^2) = \mathbb{E}(\tan^2(\Theta)) = \int_{-\theta_0}^{\theta_0} \frac{\tan^2(\theta)}{2\theta_0}\, d\theta.$$

Though you might not remember from calculus, it is easy to verify that

$$\frac{d}{d\theta}(\tan(\theta) - \theta) = \sec^2(\theta) - 1 = \tan^2(\theta)$$

and so it follows that

$$\text{Var}(Y) = \frac{1}{2\theta_0}\int_{-\theta_0}^{\theta_0} \tan^2(\theta)\, d\theta = \frac{1}{2\theta_0}(\tan(\theta) - \theta)\Big|_{-\theta_0}^{\theta_0} = \frac{\tan(\theta_0) - \theta_0}{\theta_0}.$$

Figure 5.13 illustrates this variance for $0 < \theta_0 < \pi/2$. This variance appears to grow without bound as $\theta_0 \to \pi/2$. When θ_0 is small, we can use the Maclaurin expansion $\tan(\theta_0) = \theta_0 + \theta_0^3/3 + 2\theta_0^5/15 + \cdots$ to see that

$$\text{Var}(Y) = \frac{1}{3}\theta_0^2 + \frac{2}{15}\theta_0^4 + \cdots = \frac{(2\theta_0)^2}{12} + \frac{2}{15}\theta_0^4 + \cdots.$$

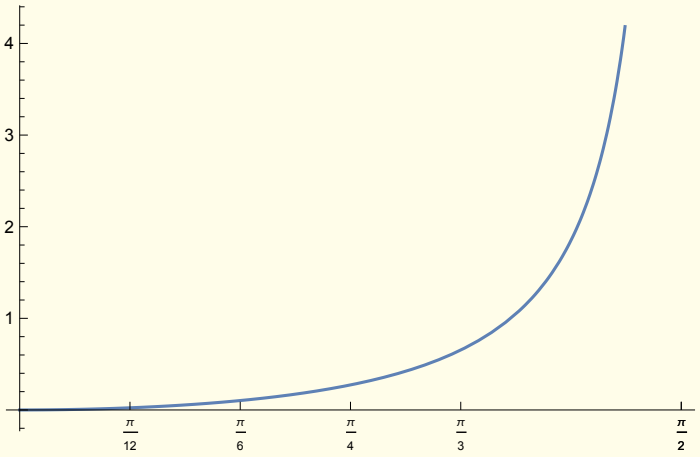


Figure 5.13: $\text{Var}(Y)$ as a function of θ_0

Upon noting that $y_0 = \tan(\theta_0) \approx \theta_0$ when θ_0 is small, this is incredibly satisfying for it says that says that the variance, to the first order, agrees with that of the uniform distribution on a small enough interval. Indeed, we expect that for a small enough angle, the sensor sees something that is approximately uniform. Just knowing the mean and variance isn't enough – our goal is to find the density of Y.

To this end, we first note that tan maps the interval $[-\theta_0, \theta_0]$ bijectively (one-to-one and onto) onto the interval $[-y_0, y_0]$ where $y_0 = \tan(\theta_0)$. With this, observe that

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\tan(\Theta) \leq y) = \mathbb{P}(\Theta \leq \tan^{-1}(y)) = \int_{-\infty}^{\tan^{-1}(y)} f_\Theta(\theta)\, d\theta$$

Now, for $y < -y_0$, $\theta = \tan^{-1}(y) < -\theta_0$ and therefore $F_Y(y) = 0$ because $f_\Theta(\theta) = 0$ for $\theta < -\theta_0$. For $-y_0 \leq y < y_0$, $-\theta_0 \leq \theta = \tan^{-1}(y) < \theta_0$ and so

$$F_Y(y) = \int_{-\theta_0}^{\tan^{-1}(y)} \frac{1}{2\theta_0}\, d\theta = \frac{\tan^{-1}(y) + \theta_0}{2\theta_0}$$

and, for $y \geq y_0$, we find $F_Y(y) = 1$. Putting everything together gives

$$F_Y(y) = \begin{cases} 0 & y < -y_0 \\ \dfrac{\tan^{-1}(y) + \theta_0}{2\theta_0} & -y_0 \leq y < y_0 \\ 1 & y_0 \leq y \end{cases}.$$

Since $\frac{d}{dy}\tan^{-1}(y) = \frac{1}{1+y^2}$, we find that

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} \frac{1}{2\theta_0(1+y^2)} & -y_0 < y < y_0 \\ 0 & \text{otherwise} \end{cases}.$$

Figure 5.14 illustrates the CDF of a PDF of Y for various values of θ_0 (equivalently y_0).

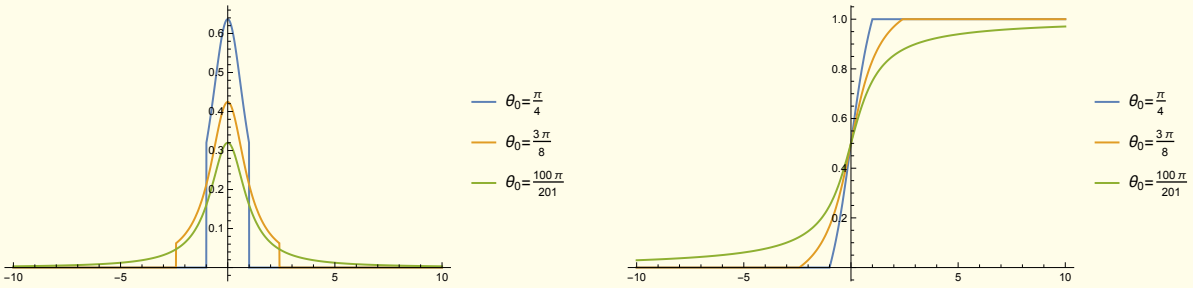


Figure 5.14: Density and Distribution Functions of Y

In the two preceding examples, we've strongly used the fact that the functions $r \mapsto \pi r^2$ and $\theta \mapsto \tan(\theta)$ were differentiable and map their domains ($[0,1]$ and $[-\theta_0, \theta_0]$) bijectively onto their ranges ($([0,\pi]$) and $[-y_0, y_0]$). Unsurprisingly, we can make this into a general result stated only in terms of probability density functions.

Proposition 5.47. *Let X be a continuous random variable with density f_X. Suppose that $\varphi : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function and is one-to-one and onto its range, $R(X) = \{x : f_X(x) \neq 0\}$. Then the random variable $Y = \varphi(X)$ is a continuous random variable with density given by*

$$f_Y(y) = f_X(\varphi^{-1}(y))\left|\frac{d}{dy}\varphi^{-1}(y)\right| = \frac{f_X(x)}{|\varphi'(x)|}$$

for (almost) every $y = \varphi(x)$ where $\varphi^{-1} : R(Y) \to R(X)$ is the necessarily differentiable inverse function of φ.

Proof. Given our assumption that φ is one-to-one on $R(X)$ and $Y = \varphi(X)$, it is clear that φ maps $R(X)$ bijectively onto $R(Y)$ with inverse φ^{-1}. Also, thanks to the continuous differentiability of φ, the in inverse function theorem guarantees

$$\frac{d}{dy}\varphi^{-1}(y) = \frac{1}{\varphi'(x)}$$

for all $y = \varphi(x)$ for which $\varphi'(x) \neq 0$. In the case that φ is strictly increasing, φ^{-1} is necessarily increasing and so

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(\varphi(X) \leq y) \leq \mathbb{P}(X \leq \varphi^{-1}(y)) = \int_{-\infty}^{\varphi^{-1}(y)} f_X(u)\,du$$

for all $y \in \mathbb{R}$. Thus, by Theorem 5.45, Y is a continuous random variable and, by virtue of the fundamental theorem of calculus and the chain rule, we have

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\int_{-\infty}^{\varphi^{-1}(y)} f_X(x)\,dx = f_X(\varphi^{-1}(y))\frac{d}{dy}\varphi^{-1}(y)$$

for almost every $y \in \mathbb{R}$. Now, under our assumption that φ is increasing, φ^{-1} is necessarily increasing and so its derivative is non-negative. Hence

$$f_Y(y) = f_X(\varphi^{-1}(y)) \left| \frac{d}{dy}\varphi^{-1}(y) \right| = \frac{f_X(x)}{|\varphi'(x)|}.$$

On the other hand, in the case that φ is decreasing, we have

$$F_Y(y) = \mathbb{P}(\varphi(X) \leq y) = \mathbb{P}(\varphi^{-1}(y) \leq X) = \int_{\varphi^{-1}(y)}^{\infty} f_X(u)\, du$$

and therefore Y is a continuous random variable with

$$f_Y(y) = -f_X(\varphi^{-1}(y)) \frac{d}{dy}\varphi^{-1}(y) = f_X(\varphi^{-1}(y)) \left(-\frac{d}{dy}\varphi^{-1}(y) \right);$$

the negative sign comes from the fact that the $\varphi^{-1}(y)$ appears in the lower limit of the integral in the penultimate equation. Since φ is decreasing, φ^{-1} must also be decreasing and so its derivative is non-positive, i.e., $-\frac{d}{dy}\varphi^{-1}(y) = \left| \frac{d}{dy}\varphi^{-1}(y) \right|$. From this, the result follows immediately. □

Example 5.37: Revisiting Examples 5.4.1 and 5.4.1

In our Pizza example, $\varphi(r) = \pi r^2$ and therefore $\varphi^{-1}(a) = \sqrt{\frac{a}{\pi}}$ for $0 \leq a \leq \pi$. Consequently,

$$\frac{d}{da}\varphi^{-1}(a) = \frac{1}{2\sqrt{\pi a}} > 0$$

for $0 < a < \pi$. Since R is uniformly distributed on $[0, 1]$, $f_R(r) = f_R(\varphi^{-1}(a)) = f_R(\sqrt{a/\pi}) = 1$ precisely when $0 \leq a \leq \pi$ and therefore

$$f_A(y) = f_R(\varphi^{-1}(a)) \frac{d}{da}\varphi^{-1}(a) = \begin{cases} \frac{1}{2\sqrt{\pi a}} & 0 < a < \pi \\ 0 & \text{else} \end{cases}.$$

This is precisely what we had found previously.

Our work above made strong use the bijectivity of the map $x \mapsto \varphi(x)$. When this isn't bijective, one must work a little bit harder – though the results can be very satisfying as you will find in the next exercise.

Exercise 5.19: Cannon Balls on a Battlefield

Suppose that a cannon is placed on a battle field at a position $x = 0$ and fires cannonballs at an angle θ from the horizontal x-axis (for $0 \leq \theta \leq \pi$) and a speed $v > 0$ (measured in meters per second). If the cannonballs of mass m are subject only to a constant gravitational force $-mg$ (measured in Newtons N) and we neglect wind resistance (and other effects such as the Coriolis force), it is straightforward to show that the cannonball will land at position

$$x = \frac{v}{g} \sin(2\theta)$$

which is measured in meters. For simplicity, let us assume that v/g is precisely one kilometer and so

$$x = \sin(2\theta)$$

measured now in kilometers. Let's now assume that the cannonballs are fired at random angles $\theta = \Theta$ where Θ is a continuous random variable distributed uniformly on $[0, \pi]$. It follows that the landing position $x = X$

becomes a continuous random variable given by

$$X = \sin(2\Theta).$$

1. What is the range of the random variable X?
2. Compute the cumulative distribution function F_X of X. Note: This is the most difficult part of this problem. To do it, you will want to invert the sine function, however, $\theta \mapsto \sin(2\theta)$ is not invertible as a function from $[0, \pi]$ to $[-1, 1]$. In fact, for every non-zero $x \in [-1, 1]$ there are exactly two values of θ for which $\sin(2\theta) = x$ (think about why this makes sense in terms of a cannonball). If you can write them down, solving this problem becomes straightforward.
3. Determine the density function f_X of X and verify that
$$1 = \int_{-\infty}^{\infty} f_X(x)\, dx$$
4. What is the expected landing position of a cannonball, $\mu = E(X)$?
5. If lots of cannonballs are fired according to this uniformly distributed angle Θ and, for an interval I, we interpret the probability $\mathbb{P}(X \in I)$ to measure the proportion of total cannonballs fired landing in I, find an interval I which is centered at μ for which one could expect to find half of the cannonballs fired.

Uniform Gives Everything

Proposition 5.48. *Let $F : \mathbb{R} \to [0, 1]$ which is non-decreasing and has $\lim_{y\to-\infty} F(y) = 0$ and $\lim_{y\to\infty} F(y) = 1$. Then there is a random variable Y with $F_Y = F$.*

Chapter 6

Multiple Random Variables

In this chapter, we will consider the situation in which we have multiple random variables defined on a single sample space Ω. For the most part, we will be concerned with two random variables and their interaction. In this, we will ask questions like: What does one random variable tell us, if anything, about the other?

6.1 Jointly Distributed Random Variables

Consider two random variables X and Y defined on a common sample space Ω. With our goal of understanding probabilities associated to events defined in terms of X and Y, let's start by talking about events associated to X and Y. In the case that X and Y are both discrete with countable ranges $R(X)$ and $R(Y)$ (which is necessarily the case when Ω is countable), anything we could ask (probabilistically) about X and Y can be answered, in principle, from understanding events of the form

$$\{X = x, Y = y\} = \{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\} = \{X = x\} \cap \{Y = y\}$$

for $x \in R(X)$ and $y \in R(Y)$. Let's take Ω to be equipped with a probability measure $\mathbb{P}$. Given that everything about a single discrete random variable can be answered in terms of its probability mass function, it makes sense make the following definition.

Definition 6.1. *Given discrete random variables X and Y defined on a common sample space Ω which is equipped with probability measure $\mathbb{P}$, the joint probability mass function associated to X and Y is the function*

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

In view of Proposition 5.19, it is natural to ask: What properties must a joint probability mass function satisfy? The answer is given by the following (unsurprising) proposition.

Proposition 6.2. *Let $p_{X,Y}$ be the probability mass function associated to two discrete random variables X and Y on a common sample space Ω. Then $p_{X,Y}$ must satisfy the following two properties:*

1. *$p_{X,Y} \geq 0$ and $p_{X,Y}(x, y) > 0$ only for $x \in R(X)$ and $y \in R(Y)$.*
2. $$\sum_{x \in R(X), y \in R(Y)} p_{X,Y}(x, y) = 1;$$

here $R(X)$ and $R(Y)$ are the ranges of the random variables X and Y respectively. Conversely, for any function $p : \mathbb{R}^2 \to \mathbb{R}$ which satisfies the above two properties and is non-zero on, at most, a countable subset of $\mathbb{R}^2$, then p is the probability mass function associated to two discrete random variables X and Y, i.e., $p = p_{X,Y}$.

It is a good exercise to prove (at least the forward direction of) the proposition above and so I'll leave that to you. Still, I think it's worth making two remarks about the proposition.

Remark 6.3. It isn't necessarily true that $p_{X,Y}(x,y) > 0$ if and only if $x \in R(X)$ and $y \in R(Y)$. For, when the random variables are related in some substantive way (think "not independent"), it is possible that the probability of the event $\{X = x, Y = y\}$ is zero when the separate events have positive probability; this is illustrated in the example below.

Remark 6.4. The summation in the second item of the proposition is a double summation. Thanks to the Fubini-Tonelli theorem, it can be computed as an iterated summation in either order as follows

$$\sum_{x\in R(X), y\in R(Y)} p_{X,Y}(x,y) = \sum_{x\in R(X)} \sum_{y\in R(Y)} p_{X,Y}(x,y) = \sum_{y\in R(Y)} \sum_{x\in R(X)} p_{X,Y}(x,y) = 1.$$

Here, the middle term means "sum over y first and then sum the result in x". The last term reverses this order.

Let $p_{X,Y}$ be a probability mass function associated to two discrete random variables X and Y and observe that, for $x \in \mathbb{R}$,

$$\{X = x\} = \bigcup_{y\in R(Y)} \{X = x, Y = y\}$$

where this union is disjoint. Consequently,

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y\in R(Y)} \mathbb{P}(X = x, Y = y) = \sum_{y\in R(Y)} p_{X,Y}(x,y).$$

and so the probability mass function for X can be computed directly from the joint probability mass function. The same is true for the probability mass function p_Y of Y. In this new ream of studying two random variables X and Y simultaneously, it is commonplace to call the probability mass functions p_X and p_Y associated to X and Y, respectively, **marginal probability mass functions**. In this language, we have:

Proposition 6.5. *Let X and Y be discrete random variables with joint probability mass function $p_{X,Y}$. Then the marginal probability mass functions p_X and p_Y are given by*

$$p_X(x) = \sum_y p_{X,Y}(x,y) \qquad \text{and} \qquad p_Y(y) = \sum_x p_{X,Y}(x,y),$$

respectively.

Example 6.1:

Consider an experiment where we roll a single perfect die (modeled by $\Omega = \{1,2,3,4,5,6\}$ and the uniform measure $\mathbb{P}$ on Ω). Consider the random variables X and Y defined by

$$X(\omega) = \begin{cases} 1 & \omega = 1 \\ 2 & \omega = 2,3,4 \\ 3 & \omega = 5,6 \end{cases} \qquad \text{and} \qquad Y(\omega) = \begin{cases} 0 & \omega \text{ odd} \\ 1 & \omega \text{ even} \end{cases}$$

for $\omega \in \Omega$. Let's compute the joint probability mass function. For $x = 1, y = 0$,

$$p_{X,Y}(1,0) = \mathbb{P}(X = 1, Y = 0) = \mathbb{P}(\omega = 1, \omega \text{ odd}) = P(\omega = 1) = \frac{1}{6}$$

Since $\omega = 1$ is not even, we have

$$p_{X,Y}(1,1) = \mathbb{P}(X = 1, Y = 1) = \mathbb{P}(\varnothing) = 0.$$

Continuing in this manner, we have

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{6} & (x,y) = (1,0) \\ 0 & (x,y) = (1,1) \\ \frac{1}{6} & (x,y) = (2,0) \\ \frac{1}{3} & (x,y) = (2,1) \\ \frac{1}{6} & (x,y) = (3,0) \\ \frac{1}{6} & (x,y) = (3,1) \end{cases} = \begin{cases} \frac{1}{6} & (x,y) = (1,0), (2,0), (3,0), (3,1) \\ \frac{1}{3} & (x,y) = (2,1) \\ 0 & (x,y) \text{ else} \end{cases}.$$

It is common to illustrate $p_{X,Y}$ in a rectangular array as follows:

1	0	$\frac{1}{3}$	$\frac{1}{6}$	
0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	
Y / X	1	2	3	

Looking at the array above, let's focus on the column for which $X = x = 1$. Summing the entries in that column, we see that

$$\frac{1}{6} + 0 = p_{X,Y}(1,0) + p_{X,Y}(1,1) = p_X(1).$$

In other words, the summation along the column for $X = x = 1$, produces the value of the marginal probability $p_X(1)$. By completely analogous computations, we find that summing along any row and column produces the marginal probability associated with that row/column. For this reason, it is customary to enter the values of p_X and p_Y along the array's margins and it is for this reason that we refer to these probabilities as *marginal* probabilities. The following array has all of this filled out.

p_X	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{3}$	
1	0	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{2}$
0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$
Y / X	1	2	3	p_Y

Having all of this information by hand allows us to answer many questions regarding the random variables X and Y. For instance, I could ask: What is the probability that $X \geq 2$ and $Y = 0$. This is

$$\mathbb{P}(X \geq 2, Y = 0) = \sum_{x=2,3} p_{X,Y}(x,0) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

and is, of course, found by summing the $(x,y) = (2,0)$ and $(3,0)$ entries of the array.

Example 6.2: Two fair coin flips

Consider an experiment where two fair coins are flipped independently (so that all outcomes in the sample space $\Omega = \{HH, HT, TH, TT\}$ are equally likely). Consider the random variable X_1 which assigns the value 1 if the first coin is heads and 0 if the first coin is tails. Similarly, let X_2 be the random variable which takes the value 1 if the second coin lands on heads and 0 if the second coin lands on tails. We can illustrate the joint probability mass function p_{X_1,X_2} and the marginal probabilities in the following table.

p_{X_1}	$\frac{1}{2}$	$\frac{1}{2}$	
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$
X_2 / X_1	0	1	p_{X_2}

Exercise 6.1: Two types of customers

A bank opens at 10:00AM and serves. primarily, two kinds of customers. The first group of customers have business accounts and the second have personal accounts. The bank would like to know how many of each type of customer to expect between the hours of 10:00AM and 11:00AM so that they are able to have a sufficient number of appropriately-trained tellers on duty. Let's denote by X and Y the number of customers with business accounts and personal accounts, respectively, that arrive at the bank between 10:00AM and 11:00AM. Though much statistical analysis, the bank finds that it is able to model these random variables by a joint probability mass function

$$p_{X,Y}(x,y) \begin{cases} \frac{e^{-(5/2)}2^x 2^{-y}}{x!y!} & x, y \in \mathbb{N} = \{0, 1, 2, \dots\} \\ 0 & \text{otherwise.} \end{cases}$$

Please do the following:

1. Verify that $p_{X,Y}$ satisfies the two properties of Proposition 6.2.
2. Using Proposition 6.5, compute the Marginal probability mass functions of X and Y. Do you recognize these as types of random variables you've encountered before? Please explain.
3. Compute the probability that at least 3 customers with business accounts and no more than 2 customers with personal accounts arrive at the bank between 10:00AM and 11:00AM.

Let's now turn our focus to (jointly) continuous random variables.

Definition 6.6. *Let X and Y be continuous random variables on a common sample space Ω which is equipped with a probability measure $\mathbb{P}$. We say that* ***the random variables X and Y are jointly continuous*** *if there is a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that*

$$\mathbb{P}(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x,y)\, dx\, dy$$

for every $x_1 \leq x_2$ and $y_1 \leq y_2$. In this case, we say that $f_{X,Y}$ is the ***joint probability density function*** *associated to X and Y.*

Example 6.3: Marble in a Box

Let's consider an experiment where we drop a marble in a box of dimension 4×6 (meters) and assume that any landing position of the marble (X, Y) is equally likely. Consistent with our discussion in Chapter 2, we can model the X and Y as jointly continuous random variables with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{24} & -2 \leq x \leq 2,\ -3 \leq y \leq 3 \\ 0 & \text{else.} \end{cases}$$

Let's use this to compute the probabilities of a couple of events. First, the probability that the marble lands in the box is

$$\begin{aligned} \mathbb{P}(-2 \leq X \leq 2, -3 \leq Y \leq 3) &= \int_{-3}^{3}\int_{-2}^{2} f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-3}^{3}\int_{-2}^{2} \frac{1}{24}\,dx\,dy \\ &= \int_{-3}^{3} \frac{x}{24}\Big|_{x=-2}^{x=2}\,dy \\ &= \int_{-3}^{3} \frac{4}{24}\,dy \\ &= 1. \end{aligned}$$

This is, of course, unsurprising as we expect the marble to actually land in the box. The probability that the marble lands in the rectangle $[0,1] \times [0,1] = \{(x,y) : 0 \leq x, y \leq 1\}$ is (by a similar computation)

$$\mathbb{P}(0 \leq X \leq 1, 0 \leq Y \leq 1) = \int_0^1 \int_0^1 \frac{1}{24}\,dx\,dy = \frac{1}{24}.$$

Observe that the event $\{x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\}$ is precisely the event that the ordered pair (X, Y) belongs to the rectangle $[x_1, x_2] \times [y_1, y_2]$, i.e.,

$$\{x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\} = \{(X,Y) \in [x_1, x_2] \times [y_1, y_2]\}.$$

In looking at the preceding example (and the definition of jointly continuous random variables), it would be nice to compute the probabilities of events which are more general than those described above. In the case of the marble in the box, for example, it would be nice to be able to compute the probability that the landing position (X, Y) falls within one unit of the center $(0, 0)$. More generally, we should be able to compute the probability that the marble lands in some general region $R \subseteq \mathbb{R}^2$. This is precisely what the following proposition allows us to do.

Proposition 6.7. *Let X and Y be jointly continuous random variables with joint probability density function $f_{X,Y}$. Then, given any[1] subset $R \subseteq \mathbb{R}^2$,*

$$\mathbb{P}((X,Y) \in R) = \iint_R f_{X,Y}(x,y)\,dA$$

We shall not prove the proposition as it follows (depending on your assumptions on $f_{X,y}$) from basic results in multivariable calculus. Let's use it to compute the probability that the marble lands within unit radius of the center $(0, 0)$.

[1] Lebesgue measurable

Example 6.4: More Marbles

Let (X, Y) be the landing position of the marble discussed in the preceding example. We have

$$\mathbb{P}((X,Y) \text{ lands within 1 unit of } (0,0)\ ) = \mathbb{P}((X,Y) \in D) = \iint_D f_{X,Y}(x,y)\,dA$$

where $D = \{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} \leq 1\}$ is the disk of radius 1 centered at $(0,0)$ in $\mathbb{R}^2$. We may express D as a so-called Type 1 region[a] by noting that

$$D = \{(x,y) : -\sqrt{1-x^2} \leq y \leq \sqrt{1-x^2},\ -1 \leq x \leq 1\}.$$

and therefore

$$\mathbb{P}((X,Y) \in D) = \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{24}\,dy\,dx = \int_{-1}^{1} \frac{\sqrt{1-x^2}}{12}\,dx$$

Unfortunately, finding an antiderivative of $\sqrt{1-x^2}$ is not terribly easy to do (but it is possible). By looking one up (say, in a calculus book or computational software), one finds that

$$\frac{d}{dx}\frac{\sin^{-1}(x) + x\sqrt{1-x^2}}{2} = \sqrt{1-x^2}$$

and so, by the fundamental theorem of calculus,

$$\mathbb{P}((X,Y) \in D) = \int_{-1}^{1} \frac{\sqrt{1-x^2}}{12}\,dx = \frac{1}{24}\left(\sin^{-1}(x) + x\sqrt{1-x^2}\right)\Big|_{-1}^{1} = \frac{\sin^{-1}(1) - \sin^{-1}(-1)}{24} = \frac{\pi}{24}.$$

There is, however, a much easier way to do this. We can instead convert to polar coordinates (r,θ) to see that $D = (r,\theta) : 0 \leq r \leq 1, 0 \leq \theta \leq 2\pi$ and $r\,dr\,d\theta = dA$ so that

$$\mathbb{P}((X,Y) \in D) = \iint_D \frac{1}{24}\,dA = \int_0^{2\pi}\int_0^1 \frac{r}{24}\,dr\,d\theta = \int_0^{2\pi}\left(\frac{r^2}{44}\right)\Big|_{r=0}^{r=1} d\theta = \frac{2\pi}{48} = \frac{\pi}{24}.$$

We see how much easier polar coordinates was in this setting.

[a] This is fairly standard terminology from multivariable calculus. If this isn't familiar to you, please come talk to me and I'll give you some references.

Exercise 6.2: So many marbles

In the above example, we heavily used the fact that $f_{X,Y}(x,y) = \frac{1}{24}$ for all $(x,y) \in D$ because the unit disk D lives entirely within the box: $[-2,2] \times [-3,3]$. In this exercise, you will need to be a little more careful. For the same jointly continuous random variables X and Y describing the landing position of marbles (as in the previous example), compute: $\mathbb{P}((X,Y) \in R)$ where R are the following regions.

1. $R = [1,4] \times [2,50] = \{(x,y) : 1 \leq x \leq 4, 2 \leq y \leq 50\}$.
2. $R = \{(x,y) : (x,y) \text{ are within one unit of } (2,3)\} = \{(x,y) : \sqrt{(x-2)^2 + (y-3)^2} \leq 1\}$.
3. R is a general region in $\mathbb{R}^2$. Hint: Your answer should be expressed in terms of area.

It is natural to ask, what properties must a joint density function possess. Mirroring Proposition 6.2, we have the following.

Proposition 6.8. *Let $f_{X,Y}$ be the joint probability density function associated to continuous random variables X*

and Y. Then $f_{X,Y}$ must satisfy the following two properties.

1. *$f_{X,Y}$ is non-negative.*
2. *We have*
$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dA = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1.$$
Conversely, any function f that satisfies the two preceding properties is the joint probability density function associated to jointly continuous random variables X and Y, i.e., $f = f_{X,Y}$.

Let's make a few important remarks.

Remark 6.9. Comparing Propositions 6.2 and 6.8, it is evident that nothing is being said about the positivity of $f_{X,Y}$ relative to the range of X and Y in Item 1 above. While there is a property analogous to that stated in Item 1 of Proposition 6.2 which is true for jointly continuous random variables, it is much more complicated to state and so I have decided not to include it. This is partially related to the fact that the probability of events of the form $\{X = x, Y = y\}$ are necessarily zero for jointly continuous random variables and so characterizing when $f_{X,Y} > 0$ is slightly harder to do.

Remark 6.10. Technically, in both the definition of joint probability density function and the proposition above, we should require a little more from $f_{X,Y}$ (and f) for us to be able to integrate them. For all intents and purposes pertaining to this class, piecewise continuous is more than enough and just about every density function we will see will have this property. The real technical requirement is that the functions need to be "Lebesgue measurable on $\mathbb{R}^2$".

Remark 6.11. The distinction between the double integrals in Property 2 above is the distinction between double integrals (defined in terms of limits of double Riemann sums and dA is the "area element") and the other is given in terms of iterated integrals, i.e., integrate in x first and then integrate the result in y. In fact, thanks to the Fubini-Tonelli theorem, we have

$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,dA = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dydx = 1.$$

In other words, the iterated integrals can be computed in either order.

Exercise 6.3: Are these joint density functions?

Given the following functions, determine whether or not it is possible for each to be the joint probability density function of two random variables. If it is not possible, say why. If it is possible, determine C.

1. $f(x,y) = \begin{cases} Cxy & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{else} \end{cases}$
2. $f(x,y) = \begin{cases} C\sin(x) & 0 \le x \le \pi,\ -\pi \le y \le \pi \\ 0 & \text{else} \end{cases}$
3. $f(x,y) = \begin{cases} C\sin(y) & 0 \le x \le \pi,\ -\pi \le y \le \pi \\ 0 & \text{else} \end{cases}$
4. $f(x,y) = Ce^{-(x^2+y^2)}$
5. $f(x,y) = \begin{cases} C & -1 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{else} \end{cases}$

6. $f(x,y) = \begin{cases} C & x = 0, 0 \leq y \leq 1 \\ 0 & \text{else} \end{cases}$

The following proposition mirrors Proposition 6.5 and, in particular, allows us to define the term marginal density function.

Proposition 6.12. *Let X and Y be jointly continuous random variables with joint probability density function $f_{X,Y}$. Then the probability density functions of X and Y are given by*

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dy$$

and

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dx,$$

respectively. For precisely the reasons explained for discrete random variables, we shall call f_X and f_Y ***marginal densities****.*

Proof. Given $x_1 \leq x_2$, we observe that

$$\begin{aligned} \mathbb{P}(x_1 \leq X \leq x_2) &= \mathbb{P}(x_1 \leq X \leq x_2, -\infty < Y < \infty) \\ &= \int_{x_1}^{x_2} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx \\ &= \int_{x_1}^{x_2} \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \right) dx. \end{aligned}$$

In view of the definition of continuous random variables and, in particular, the definition of the probability density function of the random variable X, the term in parentheses must coincide with the density of X, f_X. A similar argument gives the same conclusion for f_Y. □

Example 6.5: Marginals for Marbles

Let's return once again to our model of the landing position (X,Y) of a marble in a box. We assumed that X and Y are jointly continuous with joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{24} & -2 \leq x \leq 2,\ -3 \leq y \leq 3 \\ 0 & \text{otherwise .} \end{cases}$$

For the marginal density of X, we have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy.$$

Because $f_{X,Y}(x,y)$ is non-zero only when $-2 \leq x \leq 2$ and $-3 \leq y \leq 3$, we see that, for $|x| > 2$, (i.e., x is not in the rectangle $[-2,2]$),

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy = \int_{-\infty}^{\infty} 0\,dy = 0.$$

In the case that $|x| \leq 2$,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{-3} f_{X,Y}(x,y)\,dy + \int_{-3}^{3} f_{X,Y}(x,y)\,dy + \int_{3}^{\infty} f_{X,Y}(x,y)\,dy \\ &= \int_{-\infty}^{-3} 0\,dy + \int_{-3}^{2} \frac{1}{24}\,dy + \int_{3}^{\infty} 0\,dy \\ &= 0 + \frac{6}{24} + 0 \\ &= \frac{1}{4}. \end{aligned}$$

All together,

$$f_X(x) = \begin{cases} \frac{1}{4} & -2 \leq x \leq 2 \\ 0 & \text{else} \end{cases}$$

for $x \in \mathbb{R}$. You've seen this density before! This means that X is uniformly distributed on the interval $[-2,2]$ and, if we think about it a little, this makes sense because in view of our model. By analogous reasoning (which you should verify directly),

$$f_Y(y) = \begin{cases} \frac{1}{6} & -3 \leq y \leq 3 \\ 0 & \text{else} \end{cases}$$

for $y \in \mathbb{R}$. From this we conclude that $Y \sim \text{Unif}([-3,3])$.

Example 6.6: Darts

Consider a dart board of $1'$ radius placed on the wall with its center, the very center of the bull's eye at $(0,0)$. We can provide a crude model for the landing position (X,Y) of the dart by assuming that X and Y are jointly continuous random variables with

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & (x,y) \in D \\ 0 & \text{else} \end{cases}$$

for $(x,y) \in \mathbb{R}^2$ where $D = \{(x,y) : \sqrt{x^2+y^2} \leq 1\}$ is the closed unit disk in $\mathbb{R}^2$. To understand the landing position (X,Y) in this model, let's compute the probability of a couple of events. Consider the Bull's eye of the dart board, $B = \{(x,y) : \sqrt{x^2+y^2} \leq 0.01\}$, i.e., this is the circular subregion of D centered at $(0,0)$ with radius $r = 0.01'$. We have

$$\begin{aligned} \mathbb{P}(\{\text{bull's eye}\}) &= \mathbb{P}((X,Y) \in B) \\ &= \iint_B f_{X,Y}(x,y)\,dA \\ &= \iint_B \frac{1}{\pi}\,dA \end{aligned}$$

where we have used the fact that $B \subseteq D$, i.e., that the bull's eye is a subset of the dartboard and so the joint density is identically equal to $1/\pi$ there. Taking cues from our marble dropping example, this integral is easily computed by switching to polar coordinates where we see that $B = \{(r,\theta) : 0 \leq r \leq 0.01,\ 0 \leq \theta \leq 2\pi\}$. We have

$$\mathbb{P}((X,Y) \in B) = \int_0^{2\pi} \int_0^{0.01} \frac{1}{\pi} r\,dr\,\theta = (0.01)^2 = 0.0001.$$

Let's compute the probability that the dart lands on the wall and off of the board. If we model the wall by $W = \mathbb{R}^2 \setminus D = \{(x, y) : \sqrt{x^2 + y^2} > 1\}$, we have

$$\mathbb{P}(\{\text{Hits wall}\}) = \mathbb{P}((X, Y) \in W) = \iint_W f_{X,Y}(x, y)\, dA = \int_W 0\, dA = 0$$

where I have used the fact that $f_{X,Y}(x, y) = 0$ whenever $(x, y) \notin D$. In this way, this seems like a fairly unrealistic model of throwing darts; every time I play, I do hit the wall. Upon recalling that the double integral of the function 1 is simply the area of the region of interest, for any region $R \subseteq \mathbb{R}^2$, we find that

$$\mathbb{P}((X, Y) \in R) = \iint_R f_{X,Y}(x, y)\, dA = \iint_{R \cap D} \frac{1}{\pi}\, dA = \frac{\text{Area}(R \cap D)}{\pi}.$$

You should check that this formula actually gives the two preceding results (for B and W). Let's compute the marginals. We have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

for $x \in \mathbb{R}$. If x does not live on the dart board, i.e., $|x| > 1$, we see that this integral is zero. In the case that $|x| \leq 1$, this integral can be non-zero, but we must understand the range of y for which $(x, y) \in D$ for this particular x. Describing D as a so-called Type 1 region,

$$D = \{(x, y) : -\sqrt{1 - x^2} \leq y \leq \sqrt{1 - x^2}, -1 \leq x \leq 1\}$$

we see that, for our fixed x, $(x, y) \in D$ if and only if $-\sqrt{1 - x^2} \leq y \leq \sqrt{1 - x^2}$. Consequently, it is only for these y's that the integral for the marginal can be non-zero and we write

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi}\, dy = \frac{2}{\pi}\sqrt{1 - x^2}.$$

All together,

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1 - x^2} & -1 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

for $x \in \mathbb{R}$. This marginal density is illustrate in Figure 6.1. By an analogous computation, we find that

$$f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1 - y^2} & -1 \leq y \leq 1 \\ 0 & \text{else} \end{cases}$$

for $y \in \mathbb{R}$.

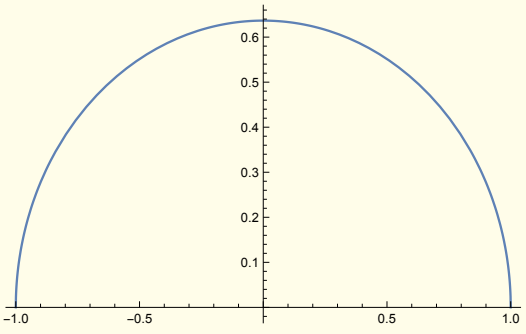


Figure 6.1: The marginal density $f_X$of X.

Exercise 6.4: A better dart model

As we saw in the preceding example, the "uniform" model for throwing darts was not terribly accurate in describing the landing position of a dart for a reasonable dart player. For example, the probability of hitting a bull's eye appeared to be unreasonably small while, simultaneously, there was no probability of hitting the wall behind the dart board – even the best dart players don't always hit the board. In this example, we model the landing position (X, Y) of a dart with jointly continuous random variables X and Y with joint density function

$$f_{X,Y}(x, y) = \frac{8}{\pi} e^{-8(x^2+y^2)}$$

for $(x, y) \in \mathbb{R}^2$. Though our dart board $D = \{(x, y) : \sqrt{x^2 + y^2} \leq 1\}$ still has radius $1'$, this density assigns positive values to points on the wall $W = \mathbb{R}^2 \setminus D$. Please do the following:

1. First, confirm that $f_{X,Y}$ is a joint probability density function.
2. Compute the probability $\mathbb{P}((X, Y) \in D)$ of a dart landing on the dart board.
3. What is the probability that the dart hits the wall, $W = \mathbb{R}^2 \setminus D$?
4. Compute the probability of getting a bull's eye, $\mathbb{P}((X, Y) \in B)$. Here, B is defined in the preceding exercise.
5. Our previous model assigned the same probability to events (within the dart board) with equal areas. Is that true for this model? Give a concrete example (or a proof) to demonstrate your conclusion.
6. Is this a more reasonable model for a "good" dart player? Explain.
7. Compute the marginal densities f_X and f_Y. Can you say anything about the random variables X and Y?

Exercise 6.5: What does Jointly Continuous Mean?

In this exercise, we think carefully about what jointly continuous means.

1. In our definition of jointly continuous random variables X and Y, we initially assumed that X and Y were continuous random variables to start with. This begs the question: If we drop the assumption that X and Y are continuous, and simply define X and Y to be jointly continuous if they have a probability density function $f_{X,Y}$, are X and Y necessarily continuous random variables? Give an argument supporting your claim or a counter example.
2. Along the lines of the above question, it is natural to wonder if there are continuous random variables X and Y (on a common sample space) which are not, together, jointly continuous. Show that, in fact, there are continuous random variables X and Y which are continuous but not jointly continuous.

6.2 Independent Random Variables

In Chapter 4, we studied independent events and argued that, if two events are independent, we can't really "learn" anything about one event from the other. Though we haven't formalized it precisely, we've "danced" around thinking about certain random variables this way. That is, though we do not yet have a definition, we should have some feeling about what it means for two random variables to be independent: Knowing something about one random variable doesn't really tell me anything about the other. Let's make this precise.

Definition 6.13. *Let X and Y be random variables on a common sample space Ω which is equipped with probability*

$\mathbb{P}$. *We say that* X *and* Y *are independent (and write* $X \perp Y$*) if, for any reasonable*[2] *subsets* I *and* J *of* $\mathbb{R}$*, the events*

$$\{X \in I\} \qquad \text{and} \qquad \{Y \in J\}$$

are independent. In other words,

$$\mathbb{P}(X \in I, Y \in J) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J) \tag{6.1}$$

for all $I, J \subseteq \mathbb{R}$.

Example 6.7: Two independent coin flips

Let's return to our example (Example 6.2) of two independent flips of a fair coin where X_1 is 1 if the first coin is heads and 0 if it is tails. Similarly, the random variable X_2 was assigned to be 1 if the second coin flip is heads and 0 if tails. Intuitively, we suspect that X_1 and X_2 are independent because X_1 depends only on the first flip and X_2 only depends on the second flip of the coin and thus, what we learn about one random variable should tell us nothing about the other. Let's confirm they are independent.

Let I and J be subsets of the real line. In the case that I contains neither 0 or 1, both of the events $\{X_1 \in I\}$ and $\{X_1 \in I,\, X_2 \in J\}$ are empty (regardless of J). Consequently,

$$\mathbb{P}(X_1 \in I, X_2 \in J) = 0 = 0 \cdot \mathbb{P}(X_2 \in J) = \mathbb{P}(X_1 \in I)\mathbb{P}(X_2 \in J).$$

Now, in the case that I contains both 0 or 1, observe that $\{X_1 \in I\} = \Omega$ and $\{X_1 \in I, X_2 \in J\} = \{X_2 \in J\}$ and therefore

$$\mathbb{P}(X_1 \in I, X_2 \in J) = \mathbb{P}(X_2 \in J) = 1 \cdot \mathbb{P}(X_2 \in J) = \mathbb{P}(\Omega)\mathbb{P}(X_2 \in J) = \mathbb{P}(X_1 \in I)\mathbb{P}(X_2 \in J).$$

Finally, let's assume that I contains either 0 or 1 so that $\mathbb{P}(X_1 \in I) = 1/2$ and, in this case, we must think about the possibilities for J. In the case that J contains neither 0 or 1, or both 0 and 1, we obtain the equation

$$\mathbb{P}(X_1 \in I, X_2 \in J) = \mathbb{P}(X_1 \in I)\mathbb{P}(X_2 \in J)$$

through precisely the same reasoning as we did for X_1 above (think it through!). In the final case that J contains either 0 or 1 (and with I also containing either 0 or 1), the intersection $\{X_1 \in I, X_2 \in J\}$ may contain only one coin flip (i.e., one outcome of the experiment) and so

$$\mathbb{P}(X_1 \in I, X_2 \in J) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(X_1 \in I)\mathbb{P}(X_2 \in J).$$

Thus, we have verified (6.1) for all possibilities of I and J and hence we may conclude that X_1 and X_2 are independent random variables.

Example 6.8: Return to Example 6.1

Let X and Y be the random variables described in Example 6.1. In contrast to the preceding two-coin example, these random variables do not appear to be independent. For example, if know that $X = 1$ so that $\omega = 1$ and so is odd, it must be the case that $Y = 0$. In other words, by learning something about one random variable, we learn something about the other. To verify concretely that these are not independent (and looking at the final table in the example), we may take $I = \{1\}$ and $J = \{1\}$ to see that

$$\mathbb{P}(X \in I, Y \in J) = \mathbb{P}(X = 1, Y = 1) = 0 \neq \frac{1}{6} \cdot \frac{1}{2} = \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J).$$

[2] You can think of I and J as being singletons $\{x\}$ and $\{y\}$ (for $x, y \in \mathbb{R}$) or intervals of real numbers. Technically, "reasonable" means Lebesgue measurable.

Although Definition 6.13 is easy to state, as you can see from the two-coin example, it is really difficult to verify because one needs to check independence of events corresponding to all subsets I and J of $\mathbb{R}$. Fortunately, when X and Y are discrete or jointly continuous, there are easy ways to study independence using probability mass/density functions. Let's focus first on the discrete setting.

Suppose that X and Y are independent random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. If X and Y are discrete, we can apply the definition of independence with $I = \{x\}$ and $J = \{y\}$ to see that

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X \in I, Y \in J) = \mathbb{P}(X \in I)\mathbb{P}(Y \in J) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all $x, y \in \mathbb{R}$. Writing this in terms of probability mass functions, we have

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

for all $x, y \in \mathbb{R}$. Thus, independent discrete random variables must have a joint mass function which is a product of their marginal probability mass functions. In fact, this property is characterizing:

Proposition 6.14. *Let X and Y be discrete random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Also, let $p_{X,Y}$, p_X, and p_Y be the associated probability mass functions to X and Y. The random variables X and Y are independent if and only if*

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

for all real numbers x and y.

Proof. In the paragraph preceding the proposition, we showed that, if the random variables X and Y are independent, then it is necessary that the joint mass function is a product of the marginals. Consequently, it remains to show that this this necessary condition for independence is also sufficient. To this end, suppose that

$$p_{X,Y}(x, y) = p_X(x)p_Y(y)$$

for all $x, y \in \mathbb{R}$ and let I and J be any subsets of $\mathbb{R}$. Upon noting that

$$\mathbb{P}(X \in I, Y \in J) = \sum_{x \in I \cap R(X)} \sum_{y \in J \cap R(Y)} p_{X,Y}(x, y),$$

a property akin to Item 4 of Theorem 5.14, we have

$$\begin{aligned}
\mathbb{P}(X \in I, Y \in J) &= \sum_{x \in I \cap R(X)} \sum_{y \in J \cap R(Y)} p_X(x)p_Y(y) \\
&= \sum_{x \in I \cap R(X)} p_X(x) \left(\sum_{y \in J \cap R(Y)} p_Y(y) \right) \\
&= \sum_{x \in I \cap R(X)} p_X(x)\mathbb{P}(Y \in J) \\
&= \mathbb{P}(Y \in J)\mathbb{P}(X \in I) \\
&= \mathbb{P}(X \in I)\mathbb{P}(Y \in J)
\end{aligned}$$

where we have made use of Item 4 of Theorem 5.14. □

Exercise 6.6: Independent or not?

Use the preceding proposition to show that the random variables in the "Two types of customers" exercise of the previous subsection are independent.

note here

Let us now turn our attention to jointly continuous random variables. We have the following result.

Proposition 6.15. *Let X and Y be jointly continuous random variables with joint density function $f_{X,Y}$ and marginal densities f_X and f_Y. Then X and Y are independent if an only if*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for almost every $x, y \in \mathbb{R}$.

note

Example 6.9: Revisiting Darts

In Example 6.6, we saw that our uniform dart model had

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \sqrt{x^2+y^2} \leq 1 \\ 0 & \text{else} \end{cases}$$

and marginal densities

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & -1 \leq x \leq 1 \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1-y^2} & -1 \leq y \leq 1 \\ 0 & \text{else} \end{cases}.$$

Here, we recall that the random variables X and Y gave the position (X, Y) of a dart thrown at a dart board of $1'$ radius where the dart player was not great (but good enough to stay on the board). We ask: Are X and Y independent? In view of Proposition 6.15, we conclude that they are not independent because $f_{X,Y}$ is clearly not a product of f_X and f_Y. If we think a little, this does seem intuitively reasonable. For example, if we know that $X > 1/2$, then the landing position Y must be somewhat limited to stay on the board (must fall between $-\sqrt{3}/2$ and $\sqrt{3}/2$ and so the full range from -1 to 1 is not accessible to Y.

The following exercise outlines a proof for proposition 6.15.

Exercise 6.7: Proof of the proposition

We shall assume that $f_{X,Y}(x,y)$, $f_X(x)$, and $f_Y(y)$ are continuous everywhere. Please do the following

1. Assume that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$ and show that X and Y are independent.
2. Assume conversely that X and Y are independent. Deduce that

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \tag{6.2}$$

   for all $x, y \in \mathbb{R}$ where $F_{X,Y}$ is the so-called joint cumulative distribution function and is given by

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$$

   for $x, y \in \mathbb{R}$.

3. Show that

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)\,du\,dv.$$

   for all $x, y \in \mathbb{R}$ and use it (and the continuity of $f_{X,Y}$) to conclude that

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = f_{X,Y}(x,y)$$

   for all $x, y \in \mathbb{R}^2$.

4. Use the above result and (6.2) to conclude that the joint probability density function is the product of marginals.

In doing all of the steps above, you have proved the proposition in the case that the densities are continuous. It takes a little more work to prove the result in general (and a result called the Lebesgue differentiation theorem), but it works though the conclusion must be restricted to "almost every".

Example 6.10: Buffon's Needle

One of the oldest problems in probability theory (in fact, the original problem of geometric probability) is the problem of Buffon's needle. This problem, posed in the late eighteenth century by Georges-Louis Leclerc, Comte de Buffon, asks: If a needle (or pin) of length l is dropped on a wooden floor made of parallel boards of width d, what is the probability that the needle with lie across the strip between two boards? Figure 6.2 illustrates this situation where $1,500$ needles (with $d = l$) have been dropped onto a floor; these randomly oriented needles have been colored green if they lie across (or intersect) a strip and red if they do not.

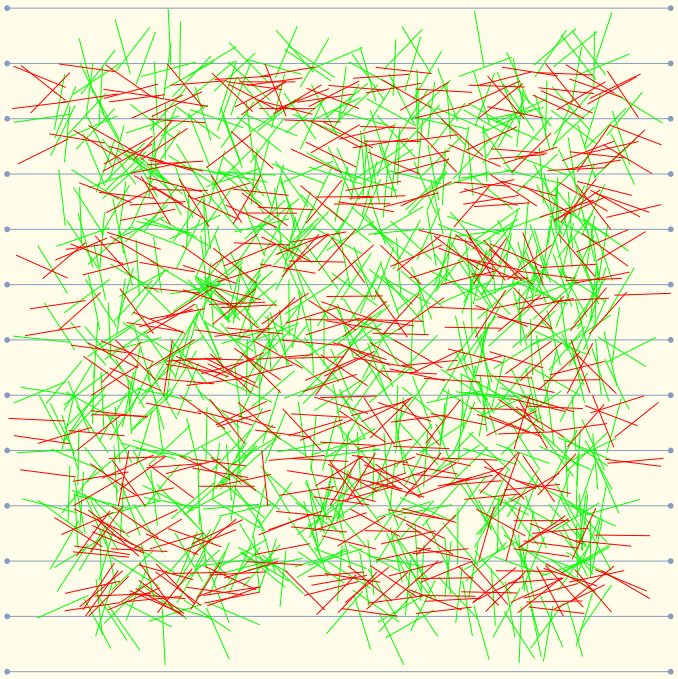
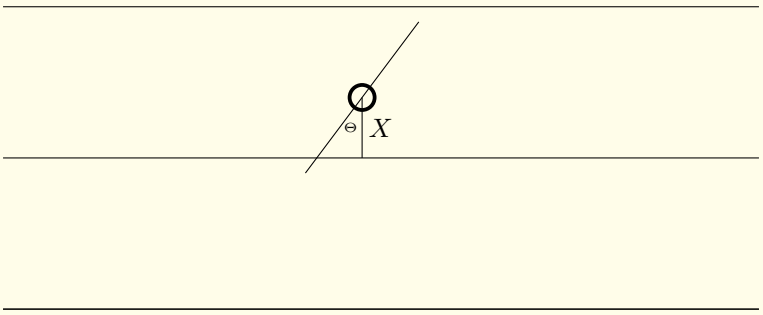


Figure 6.2: An illustration of needles on a floor ($d = l$)

We will solve the problem of Buffon's needle in the so-called "short needle" case, i.e, the case in which $l \leq d$. As the event that a given needle lands on a line depends on the needle's landing position and (angular) orientation, this is very naturally a two parameter problem. Let's describe two such parameters which will allow us to solve the problem. Let us denote the distance from the center of the needle to the nearest horizontal line (strip) by X and let Θ denote the acute angle that the needle makes with the (shortest) line segment from the needle's center to the nearest horizontal line.



Given that the hypotenuse of half of the needle is $l/2$, we see that the needle intersects the strip if and only if

$$X \leq \frac{l}{2}\cos(\Theta).$$

In looking at the phrasing of the initial problem, our goal is to compute

$$\mathbb{P}(\text{Intersect}) = \mathbb{P}\left(X \leq \frac{l}{2}\cos(\Theta)\right) = \mathbb{P}((X,\Theta) \in R)$$

where

$$R = \left\{(x,\theta) : x \leq \frac{l}{2}\cos(\theta)\right\}.$$

In thinking about the process (dropping the needle and it landing at a random location making a random acute angle with the perpendicular), it is reasonable to assume that X and Θ are both uniformly distributed random variables with

$$X \sim \text{Unif}([0, d/2]) \qquad \text{and} \qquad \Theta \sim \text{Unif}([0, \pi/2]).$$

We note that X, being the distance from the center of the needle to the nearest line, must have $0 \leq X \leq d/2$ and Θ, being an acute angle, must have $0 \leq \Theta \leq \pi/2$. Therefore, we have the marginal densities

$$f_X(x) = \begin{cases} \frac{2}{d} & 0 \leq x \leq \frac{d}{2} \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad f_\Theta(\theta) = \begin{cases} \frac{2}{\pi} & 0 \leq \theta \leq \frac{\pi}{2} \\ 0 & \text{else} \end{cases}.$$

But, what about the joint density? Thinking about the situation for a moment, it is reasonable to assume that X is not affected by Θ and vice versa. In other words, it is reasonable to assume that X and Θ are independent random variables. Thus, in view of Proposition 6.15, the joint probability density for X and Θ is

$$f_{X,\Theta}(x,\theta) = f_X(x)f_\Theta(\theta) = \begin{cases} \frac{4}{\pi d} & 0 \leq x \leq \frac{d}{2},\ 0 \leq \theta \leq \frac{\pi}{2} \\ 0 & \text{else} \end{cases}$$

for $(x, \theta) \in \mathbb{R}^2$. With this and in view of Proposition 6.7, we have

$$\begin{aligned}
\mathbb{P}(\text{Intersect}) &= \mathbb{P}((X, \Theta) \in R) \\
&= \iint_R f_{X,\Theta}(x, \theta)\, dx\, d\theta \\
&= \int_0^{\pi/2} \int_0^{l\cos(\theta)/2} \frac{4}{\pi d}\, dx\, d\theta \\
&= \int_0^{\pi/2} \frac{2l\cos(\theta)}{\pi d}\, d\theta \\
&= \frac{2}{\pi d}\sin(\theta)\Big|_0^{\pi/2} \\
&= \frac{2l}{\pi d}
\end{aligned}$$

where we have used the short-needle condition that $l \leq d$

In thinking about this problem from the beginning, this problem seems to have nothing to do with π – we are just asking about the probability that a needle intersects a horizontal line. In the end, we find that

$$\pi = \frac{2l}{d}\frac{1}{\mathbb{P}(\text{Intersect})}$$

In frequentist interpretation of probability, an interpretation which is justified by the forthcoming law of large numbers, if we toss n needles at the floor and denote by $I(n)$ the number of needles which intersect the line, then

$$\mathbb{P}(\text{Intersect}\}) = \lim_{n\to\infty} \frac{I(n)}{n}.$$

Correspondingly,

$$\pi = \lim_{n\to\infty} \frac{2ln}{dI(n)}.$$

This gives us an amazing way to approximate the value of π! In fact, with the $n = 1,500$ needles tossed in the simulation illustrated in Figure 6.2 (wherein $d = l$), $I(n) = 941$ and so

$$\frac{2ln}{dI(n)} = \frac{2(1500)}{941} = 3.1880977683315623$$

which is certainly not a bad approximation for π. It is certainly closer to the actual value of π than that proposed by the Indiana State House of Representatives in the so-called House Bill 246 of 1897.

Exercise 6.8: More of Buffon

This exercise treats two variations of the problem of Buffon's needle.

1. In looking at the preceding solution, you might be bothered by the assumption that Θ is uniformly distributed between 0 and $\pi/2$, i.e., we chose the acute angle. In fact, it is somewhat more reasonable to me to ask that $\Theta \sim \text{Unif}([-\pi/2, \pi/2])$ in which case we are saying that any angle be possible (and not simply choose the acute one). Show that, in fact, it doesn't matter, i.e., $\mathbb{P}(\text{Intersect}) = (2l)/\pi d$ when $\Theta \sim \text{Unif}([-\pi/2, \pi/2])$.

2. Under the so-called "long needle" assumption, i.e., $l \geq d$, show that

$$\mathbb{P}(\text{Intersect}) = \frac{2l}{\pi d} - \frac{2}{\pi d}\left(\sqrt{l^2 - d^2} + d\sin^{-1}(d/l)\right) + 1$$

Exercise 6.9: Rectangles

Let X and Y be jointly continuous random variables with joint probability density $f_{X,Y}(x,y)$. Throughout this problem, we will assume that $f_{X,Y}$ and the marginals f_X and f_Y are all continuous functions (though this need not be the case).
We recall that the range of X, $R(X)$, can be interpreted as the set of real numbers x such that $f_X(x) > 0$, i.e.,

$$R(X) = \{x \in \mathbb{R} : f_X(x) > 0\}.$$

Similarly, the range of the ordered pair (X,Y) (which is a function from Ω into $\mathbb{R}^2$) can be interpreted as the set of pairs (x,y) for which $f_{X,Y}(x,y) > 0$. In other words,

$$R((X,Y)) = \{(x,y) \in \mathbb{R}^2 : f_{X,Y}(x,y) > 0\}.$$

1. If X and Y are independent, show that $R(X,Y)$ must be a rectangle $\mathbb{R}^2$ (specifically, $R(X,Y) = R(X) \times R(Y)$).
2. Use this to confirm (once again) that the random variables X and Y in Example 6.6 are not independent.

Note here

Now that we've considered pairs of independent random variables, it is natural to ask about the independence of larger collections collections of random variables. Thankfully, Definition 6.13 gives us the language we need to do this.

Definition 6.16. *Let $X_1, X_2, X_3, \ldots$ be a finite or countably infinite collection of random variables all defined on a common sample space Ω equipped with probability measure $\mathbb{P}$. We say that this collection of random variables is independent (and synonymously that the random variables are independent) if, for any finite collection of indices $\{j_1, j_2, \ldots, j_k\} \subseteq \mathbb{N}_+$ and any choice of subsets of real numbers $I_1, I_2, \ldots, I_k \subseteq \mathbb{R}$, the events*

$$\{X_{j_1} \in I_1\},\ \{X_{j_2} \in I_2\},\ \{X_{j_3} \in I_3\}, \qquad \text{and} \qquad \{X_{j_n} \in I_n\}$$

are independent. Equivalently, the random variables $X_1, X_2, X_3, \ldots$ are independent if, for any finite collection of indices $\{j_1, j_2, \ldots, j_k\}$ and any collection of subsets of real numbers $I_1, I_2, \ldots, I_j$, we have

$$\mathbb{P}(X_{j_1} \in I_1, X_{j_2} \in I_2, \ldots, X_{j_k} \in I_k) = \mathbb{P}(X_{j_1} \in I_1) \times \mathbb{P}(X_{j_2} \in I_2) \times \cdots \times \mathbb{P}(X_{j_k} \in I_k).$$

Now, for finite collections $X_1, X_2, \ldots, X_n$ of discrete random variables, it is easy to characterize independence with joint probability mass functions akin to Proposition 6.2. In this case, the random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$p(x_1, x_2, \ldots, x_n) := \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\cdots p_{X_n}(x_n)$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$. Here, $p(x_1, x_2, \ldots, x_n)$ can be recognized as a joint probability mass function for all of the random variables $X_1, X_2, \ldots, X_n$. For an infinite collection of discrete random variables, a analogous characterization is a bit harder to state (though it is doable). It can be done simply by demanding that probability mass functions associated to any finite subcollection satisfy the preceding property.

Example 6.11: Independent Bernoulli Random Variables

Consider an infinite sequence of independent coin flips (i.e., we flip a coin over and over again independently ad infinitum). For each $j = 1, 2, \ldots$, let X_j encode the result of the jth flip by assigning $X_j = 1$ if the jth flip is heads and 0 if it is tails. If our coin is biased with probability p of coming up heads on each flip and

$q = 1 - p$ of coming up tails on each flip, we have, for each $j = 1, 2, \ldots,$

$$p_{X_j}(x_j) = \begin{cases} p & x_j = 1 \\ q & x_j = 0 \\ 0 & x_j \text{ else} \end{cases}$$

for $x_j \in \mathbb{R}$. In other words, for each $j = 1, 2, \ldots,$ the random variable X_j is Bernoulli with $X_j \sim \text{Ber}(p)$. This is an infinite sequence of independent random variables and, because each has the same marginal distribution, we will often say that this colleciton is **independent and identically distributed** which is commonly abbreviated as "i.i.d."

We should also mention a corresponding characterization of independence for collections of continuous random variables. Given a finite collection $X_1, X_2, \ldots, X_n$ of continuous random variables on a common sample space Ω, the collection is independent if and only if, for any region $R \subseteq \mathbb{R}^n$,

$$\mathbb{P}((X_1, X_2, \ldots, X_n) \in R) = \int \int \cdots \int_R f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)\, dx_1\, dx_2 \cdots dx_n.$$

In other words, $X_1, X_2, \ldots, X_n$ have a joint probability density function f given by the product of marginals, i.e.,

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

Note

A final note is worth mentioning before we move on to the next subsection. In the spirit of Exercise 4.6, for a sample space to support many independent events (and hence many independent random variables), it must be very large. There is a fairly straightforward way to write down (construct) sample spaces Ω on which one may have finite collections of independent random variables. For example, to construct a finite collection of n independent Bernoulli random variables, we could take

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots, \omega_n) : \omega_k = H \text{ or } T \text{ for } k = 1, 2, \ldots n\} = \{H, T\}^n$$

i.e., Ω is the n-fold Cartesian product of the two-outcome sample space $\{H, T\}$. On this Ω, the probability measure $\mathbb{P}$ which assigns $\mathbb{P}(\omega) = p^k q^{n-k}$ to each event $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ having k heads and $n - k$ tails makes the collection of n random variables $X_1, X_2, \ldots, X_n$ defined by

$$X_k(\omega) = \begin{cases} 1 & \text{if } \omega_k = H \\ 0 & \text{if } \omega_k = T \end{cases}$$

an i.i.d collection of Bernoulli random variables each with parameter p. We note that $\#(\Omega) = 2^n$ and this grows without bound as n increases. Thinking about infinite collections of random variables, how could we go about constructing a sample space that would support them? More coarsely, we could ask: Do sample spaces (equipped with probability measures) exist which are "large" enough to support an infinite collection of independent random variables? Taking the above discussion as motivation, we would expect formally that $\Omega = \{H, T\}^\infty$. In other words, we expect that Ω is not just an infinite set but more explicitly one that is infinite dimensional. In this case, it isn't clear (at least to me) how one would construct and appropriate $\mathbb{P}$. In fact, it is a fairly deep theorem that such sample spaces do exist and have probability measures which support infinite collections of independent random variables. The proof of any such result must rely on some pretty deep mathematics (e.g. Tychonoff's thoerem) and, with it, the standard course of proof is to construct a so-called infinite product measure. Refer to Rao for a proof? In any case, when we later say "let $X_1, X_2, \ldots$ be a collection of independent and identically distributed random variables", you should know that we are relying on some very deep mathematics to do so.

Exercise 6.10: Three independent random variables

Suppose that A, B, and C are independent random variables which are all uniformly distributed on $[0, 1]$.

1. Find the joint continuous density function of A, B, and C, $f_{A,B,C}(a, b, c)$.
2. What is the probability that all roots of the quadratic equation
$$Ax^2 + Bx + C = 0$$
are real? Hint: This is the same as asking "what is the probability that the discriminant is positive?"

6.3 Joint Expectation

Now that we have ways to understand random multiple random variables simultaneously, we are in the position to talk about the expectation of random variables which are "built" using them. For example, if we model the landing position (X, Y) of a dart on a dart board with random variables X and Y, can we compute the expected value of the random variable
$$d = \sqrt{X^2 + Y^2}$$
which measures the distance to the center of the board? For another example, consider an experiment in thermodynamics where one takes measurements of several samples of an ideal gas (with constant volume $V = V_0$) and records the temperature T and pressure P for each sample (which we represent as random variables). Then, according to the ideal gas law, the number of particles in each sample is given by
$$N = \frac{PV_0}{\kappa_B T}$$
where κ_B is the Boltzmann constant. If know the distributions of the pressure P and temperature T of the samples, we would like to compute the average number of particles amount the samples, $\mathbb{E}(N)$, even though we likely have no information[3] about the distribution of N. Seeing N as a function of P and T (for which we have data), we should be able to compute this with knowledge of their joint distribution.

In general, suppose that X and Y are random variables on a sample space Ω equipped with probability $\mathbb{P}$ and, for a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$, consider the random variable $\varphi(X, Y)$ (note, this is the function taking $\omega \in \Omega$ to the real number $\varphi(X(\omega), Y(\omega))$. In the case that Ω is countable, Definition 5.4 allows us to compute the expectation straightforwardly:
$$\mathbb{E}(\varphi(X, Y)) = \sum_{\omega \in \Omega} \varphi(X(\omega), Y(\omega))\mathbb{P}(\omega)$$
whenever this sum/series is convergent. However, as we have discussed before, this is generally impossible to do as we rarely have complete information about the underlying probability measure. Instead, we make use of the following theorem (which does not require Ω to be countable.

Theorem 6.17. *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Given a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$, consider the random variables $\varphi(X, Y)$.*

1. *If X and Y are discrete with joint probability mass function $p_{X,Y}$, then*
$$\mathbb{E}(\varphi(X, Y)) = \sum_{x,y} \varphi(x, y)p_{X,Y}(x, y)$$
provided that this sum/series converges.

[3]Generally, N is on the order of Avogadro's number and so, at best, we can only hope for some statistical information.

2. *If X and Y are jointly continuous with joint density function $f_{X,Y}$, then*

$$\mathbb{E}(\varphi(X,Y)) = \iint_{\mathbb{R}^2} \varphi(x,y) f_{X,Y}(x,y)\, dA$$

provided that this double integral converges.

Example 6.12: Returning to Example 6.1

In Example 6.1, we considered random variables X and Y with joint probability mass function

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{6} & (x,y) = (1,0), (2,0), (3,0), (3,1) \\ \frac{1}{3} & (x,y) = (2,1) \\ 0 & \text{otherwise.} \end{cases}$$

Let's compute the expectation of $Z = XY$. By virtue of Theorem 6.17 (with $\varphi(x,y) = xy$), we have

$$\begin{aligned} \mathbb{E}(XY) &= \sum_{x,y} xy\, p_{X,Y}(x,y) \\ &= \sum_{x=1,2,3} \sum_{y=0,1} xy\, p_{X,Y}(x,y) \\ &= \sum_{x=1,2,3} ((x\cdot 0) p_{X,Y}(x,0) + (x\cdot 1) p_{X,Y}(x,1)) \\ &= \sum_{x=1,2,3} x p_{X,Y}(x,1) \\ &= 1\cdot p_{X,Y}(1,1) + 2\cdot p_{X,Y}(2,1) + 3\cdot p_{X,Y}(3,1) \\ &= 1\cdot 0 + 2\cdot \frac{1}{3} + 3\cdot \frac{1}{6} \\ &= \frac{7}{6}. \end{aligned}$$

Example 6.13: Distance to center of the dart board

Consider, once again, our dart-throwing example where the landing position (X,Y) of darts is modeled by jointly continuous random variables X and Y with

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & (x,y) \in D \\ 0 & \text{else} \end{cases}$$

where $D = \{(x,y) \in \mathbb{R}^2 : \sqrt{x^2+y^2} \leq 1\}$ is the region of $\mathbb{R}^2$ occupied by the dart board. Given $\varphi(x,y) = \sqrt{x^2+y^2}$, the distance from the center of the dart board to the landing position of the dart is

$$d = \varphi(X,Y) = \sqrt{X^2+Y^2}.$$

Thus, the expected value of d is given by

$$\mathbb{E}(d) = \iint_{\mathbb{R}^2} \varphi(x,y) f_{X,Y}(x,y)\, dA = \iint_D \sqrt{x^2+y^2}\frac{1}{\pi} dA = \frac{1}{\pi}\iint_D \sqrt{x^2+y^2}\, dx\, dy$$

where we have used the fact that $f_{X,Y}(x,y) = 0$ for all $(x,y) \in \mathbb{R}^2 \setminus D$. While this integral can be computed directly using iterated integrals (and by recognizing D as a Type 1 (or 2) region), we can more easily do the

computation by converting to polar coordinates. We recall that, in polar coordinates, $D = \{(r, \theta) : 0 \leq r \leq 1, -\pi < \theta \leq \pi\}$, $r = \sqrt{x^2 + y^2}$ and $dA = dxdy = r\, dr\, d\theta$. By the polar coordinate integration formula, we have

$$\mathbb{E}(d) = \frac{1}{\pi} \int_{-\pi}^{\pi} \int_0^1 r \cdot r\, dr\, d\theta = \frac{2\pi}{\pi} \int_0^1 r^2\, dr = \frac{2}{3}.$$

Thus, in this model, the average distance from the landing position to the center is 2/3 of entire radius of the dart board– not a terribly good dart player, I'd say!

Exercise 6.11: A better dart player?

Using the model described in Exercise 6.4 in which

$$f_{X,Y}(x, y) = \frac{8}{\pi} e^{-8(x^2+y^2)}$$

for $(x, y) \in \mathbb{R}^2$, find the expected value of $d = \sqrt{X^2 + Y^2}$. On average, does this model describe a better dart player (if the goal is to get close to the bull's eye) than the previous example?

With our new found way to compute expectation using joint probability mass functions and density functions, it is worthwhile to verify that we are still getting the same formulas as guaranteed by our big expectation theorem, Theorem 5.41, when $\varphi(X, Y)$ is just a function of X or Y alone, e.g., when $\varphi(X, Y) = X$. Let's give this a shot as a "sanity check". Suppose that X and Y are discrete random variables with joint probability mass function $p_{X,Y}(x, y)$ and let $\varphi(x, y) = x$. Then, by Theorem 6.17, we have

$$\mathbb{E}(X) = \mathbb{E}(\varphi(X, Y)) = \sum_{x,y} \varphi(x, y) p_{X,Y}(x, y) = \sum_{x,y} x \cdot p_{X,Y}(x, y).$$

If we break this into an iterated sum, we can sum over y first to find that

$$\mathbb{E}(X) = \sum_x \sum_y x \cdot p_{X,Y}(x, y) = \sum_x x \left(\sum_y p_{X,Y}(x, y) \right) = \sum_x x \cdot p_X(x)$$

where we have used the fact that the marginal probability mass function for X satisfies

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

for each $x \in \mathbb{R}$. Therefore

$$\mathbb{E}(X) = \sum_x x p_X(x)$$

which is what we had before. By similar reasoning, we find that, for any α and β,

$$\begin{aligned}
\mathbb{E}(\alpha X + \beta Y) &= \sum_{x,y} (\alpha x + \beta y) p_{X,Y}(x, y) \\
&= \alpha \sum_{x,y} x p_{X,Y}(x, y) + \beta \sum_{x,y} y p_{X,Y}(x, y) \\
&= \alpha \sum_x x \left(\sum_y p_{X,Y}(x, y) \right) + \beta \sum_y y \left(\sum_x p_{X,Y}(x, y) \right) \\
&= \alpha \sum_x x p_X(x) + \beta \sum_y y p_Y(y) \\
&= \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)
\end{aligned}$$

and so we have confirmed the linearity assertion of Theorem 5.41 using Theorem 6.17 for discrete random variables. For good practice, you should try to work through these same details in the case that X and Y are jointly continuous.

As we saw when we originally introduced the expectation for a single random variable, the mean and variance played a useful role in giving us information about the random variable. For two random variables, there is another useful number called the covariance.

Definition 6.18. *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Let X have mean μ_X and Y have mean μ_Y. The **covariance of X and Y** is the number*

$$\text{Cov}(X,Y) = \mathbb{E}((X-\mu_x)(Y-\mu_y))$$

whenever this expectation exists.

We remark that $\text{Var}(X) = \text{Cov}(X,X)$. As the variance provides a measure of the "spread" away from a random variable's mean, the covariance tends to provide a good measure of how "correlated" two random variables are. We shall soon provide a partial justification for this statement but, for now, we should work out some useful formulas for $\text{Cov}(X,Y)$ in the case of discrete or jointly continuous random variables. We have the following.

Proposition 6.19. *Let X and Y be random variables on a common sample space Ω with probability $\mathbb{P}$. Let μ_X and μ_Y denote the mean of X and Y respectively. Then the covariance of X and Y exists if and only if $\mathbb{E}(XY)$ exists and, in this case,*

$$\text{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\mu_Y.$$

In the case that X and Y are discrete with joint probability mass function $p_{X,Y}$, we have

$$\text{Cov}(X,Y) = \sum_{x,y}(x-\mu_X)(y-\mu_Y)p_{X,Y}(x,y) = \left(\sum_{x,y} xy\, p_{X,Y}(x,y)\right) - \mu_X\mu_Y.$$

In the case that X and Y are jointly continuous with joint probability density $f_{X,Y}$, we have

$$\text{Cov}(X,Y) = \iint_{\mathbb{R}^2}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)\,dA = \iint_{\mathbb{R}^2} xy\, f_{X,Y}(x,y)\,dA - \mu_X\mu_Y$$

Proof. We have

$$\text{Cov}(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)) = \mathbb{E}(XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y).$$

Using the linearity of expectation (Theorem 5.41), we have

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}(XY) - \mu_Y\mathbb{E}(X) - \mu_X\mathbb{E}(Y) + \mu_X\mu_Y\mathbb{E}(1)\\ &= \mathbb{E}(XY) - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y = \mathbb{E}(XY) - \mu_X\mu_Y.\end{aligned}$$

where we have used the fact that $\mathbb{E}(1) = 1$ and noted that $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$. From this, the remaining formulas (for discrete and jointly continuous random variables) now follow directly from Theorem 6.17 using $\varphi(x,y) = (x-\mu_X)(y-\mu_Y)$ and $\varphi(x,y) = xy$. □

Example 6.14: Two simple examples of covariance

In Example 6.1 (and 6.12), we considered two discrete random variables X and Y with joint probability mass function

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{6} & (x,y) = (1,0),(2,0),(3,0),(3,1) \\ \frac{1}{3} & (x,y) = (2,1) \\ 0 & (x,y) = (0,0) \end{cases}.$$

Using the marginal distributions which we found in Example 6.1, we can easily compute $\mu_X = \mathbb{E}(X) = 1(1/6) + 2(1/2) + 3(1/3) = 13/6$ and $\mu_Y = \mathbb{E}(Y) = 0(1/2) + 1(1/2) = 1/2$. As we computed $\mathbb{E}(XY) = 7/6$

in Example 6.12, we have

$$\text{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\mu_Y = \frac{7}{6} - \left(\frac{13}{6}\right)\left(\frac{1}{2}\right) = \frac{1}{12}.$$

In Example 6.2 (and the latter independence one), we looked at two independent coin flips of a fair coin. In that example, X_1 took the value 1 if the first flip was heads and 0 if it was tails; X_2 had the same assignment based on the result of the second coin flip. We found that

$$p_{X_1,X_2}(x,y) = \begin{cases} \frac{1}{4} & (x_1,x_2) = (0,0),(1,0),(0,1),(1,1) \\ 0 & \text{otherwise.} \end{cases}$$

In this case, we can directly compute $\mu_{X_1} = \mu_{X_2} = \frac{1}{2}$ and

$$\begin{aligned} \text{Cov}(X_1,X_2) &= \mathbb{E}(X_1X_2) - \mu_{X_1}\mu_{X_2} \\ &= \sum_{x_1,x_2} x_1x_2\, p_{X_1,X_2}(x_1,x_2) - \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) \\ &= (0\cdot 0)p_{X_1,X_2}(0,0) + (1\cdot 0)p_{X_1,X_2}(1,0) + (0\cdot 1)p_{X_1,X_2}(0,1) + (1\cdot 1)p_{X_1,X_2}(1,1)) - \frac{1}{4} \\ &= \frac{1}{4} - \frac{1}{4} \\ &= 0. \end{aligned}$$

Exercise 6.12: Covariance of two types of customers

We recall Example 6.1 in which X and Y represent the number of customers with business accounts and personal accounts that arrive at a bank between 10:00AM and 11:00AM, respectively. In this case, our joint probability mass function is

$$p_{X,Y}(x,y) = \begin{cases} \frac{e^{-(5/2)}2^x2^{-y}}{x!y!} & x,y \in \mathbb{N} \\ 0 & \text{else.} \end{cases}$$

Please do the following:

1. Compute $\mu_X = \mathbb{E}(X)$ and $\mu_Y = E(Y)$. You can do this directly or, in the case that you recognized the distributions of X and Y, you can appeal to their well-known means.
2. Compute $\mathbb{E}(XY)$.
3. Compute $\text{Cov}(X,Y)$.

Exercise 6.13: Properties of Covariance

Let X, Y and Z be random variables. Use the linearity of expectation to show that following:

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.
2. For any constants α and β, $\text{Cov}(\alpha X, \beta Y) = \alpha\beta\,\text{Cov}(X,Y)$.
3. $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$ and $\text{Cov}(X,Y+Z) = \text{Cov}(X,Y) + \text{Cov}(X,Z)$.

You may assume that all expectations exists (i.e., you don't need to worry about absolute convergence). In mathematical terms, the above shows that the covariance is a "bilinear form" and, if interpreted correctly,

it gives a way to do geometry on the set/space of random variables in the same way that the dot product allows us to study geometry (e.g., measure angles, sizes, etc.) in Euclidean space.

We complete this section by studying joint expectation in the case that random variables X and Y are continuous. We have the following result.

Proposition 6.20. *Let X and Y be independent random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Then, for any single-variable real-valued functions g and h , we have*

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

whenever the expectations $\mathbb{E}(g(X))$ and $\mathbb{E}(h(Y))$ exists. In particular,

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

Proof. I will prove the proposition under the assumption that X and Y are jointly continuous and encourage you to work out the details in the discrete case. Since $X \perp Y$, Proposition 6.15 guarantees that

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $x, y \in \mathbb{R}$. Thus,

$$\begin{aligned}
\mathbb{E}(g(X)h(Y)) &= \iint_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)\,dA \\
&= \iint_{\mathbb{R}^2} g(x)h(y)f_X(x)f_Y(y)\,dA \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (h(y)f_Y(y))(g(x)f_X(x))\,dx\,dy \\
&= \int_{-\infty}^{\infty} h(y)f_Y(y)\left(\int_{-\infty}^{\infty} g(x)f_X(x)\,dx\right)dy \\
&= \int_{-\infty}^{\infty} h(y)f_Y(y)\mathbb{E}(g(X))\,dy \\
&= \mathbb{E}(g(X))\int_{-\infty}^{\infty} h(y)f_Y(y)\,dy \\
&= \mathbb{E}(g(X))\mathbb{E}(h(Y))
\end{aligned}$$

where we have used the Fubini-Tonelli theorem and the linearity of the integral. □

As a consequence of the theorem, observe that

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X\mu_Y = \mathbb{E}(X)\mathbb{E}(Y) - \mu_X\mu_Y$$

whenever $X \perp Y$. Since $\mathbb{E}(X) = \mu_X$ and $\mathbb{E}(Y) = \mu_Y$, we have established:

Corollary 6.21. *If $X \perp Y$, then $\mathrm{Cov}(X,Y) = 0$.*

In looking back to Example 6.7, we saw that the random variables X_1 and X_2 (given initially in Example 6.2) are independent and so $\mathrm{Cov}(X_1, X_2) = 0$; of course, we had established this explicitly in Example 6.14. In Example 6.14, we also showed that $\mathrm{Cov}(X,Y) = 1/6$ for the random variables appearing in Example 6.1 and so, in view of the above corollary, we may conclude that X and Y are not independent. Of course, this is consistent with the conclusion we made in Example 6.8. In view of the above corollary, is the result you found in Exercise 6.11 surprising?

6.4 Conditioning on random variables and conditional expectation

When we introduced conditional probability, we discussed that process of conditioning was really one of the transference of information. That is, for events A and B, the probability of A given B, $\mathbb{P}(A|B)$, tells us the likelihood of A if we know that the event B happened (or will happen). This gives us a way to update our predictions of events as we learn information. In this section, we shall study these same ideas in the context of random variables. In particular, we will learn how to update our predictions involving one random variable if we know something about another.

In this section, we will consider a pair of random variables X and Y which are either discrete or jointly continuous. The general study of conditioning on random variables is one that involves some fairly sophisticated mathematical machinery (notions of sigma algebras, filtrations, and projection operators, to name a few) and so we will avoid a more general discussion with the hope that you will see it in a future presentation of this subject.

Definition 6.22. *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$.*

1. *If X and Y are discrete with joint probability mass function $p_{X,Y}$, the conditional probability mass function of X given that $Y = y$ is defined by*

$$p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

for $x \in \mathbb{R}$ and y such that $p_Y(y) > 0$ (equivalently, for $y \in R(Y)$.) Here, of course, p_Y is the marginal probability mass function of Y.

2. *In the case that X and Y are jointly continuous with joint probability density function $f_{X,Y}(x, y)$, the conditional probability density function of X given that $Y = y$ is defined by*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

for any real numbers x and y such that $f_Y(y) > 0$.

Example 6.15: Return to Example 6.1

In view of Example 6.1, one that we now understand well, let's discuss the conditional probability mass function $p_{X|Y}$. We recall that

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{6} & (x, y) = (1, 0), (2, 0), (3, 0), (3, 1) \\ \frac{1}{3} & (x, y) = (2, 1) \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad p_Y(y) = \begin{cases} \frac{1}{2} & y = 0, 1 \\ 0 & \text{else} \end{cases}.$$

For $x = 1$ and $y = 0$, we have

$$p_{X|Y}(1|0) = \mathbb{P}(X = 1|Y = 0) = \frac{p_{X,Y}(1, 0)}{p_Y(0)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

Since $p_X(1) = \mathbb{P}(X = 1) = \frac{1}{6}$, we can interpret this conditional probability $p_{X|Y}(1|0) = 1/3$ by saying that it is twice as likely that $X = 1$ if we know that $Y = 0$. Carrying out this computation for all (relevant) x and y, we find

$$p_{X|Y}(x|0) = \begin{cases} \frac{1}{3} & x = 1, 2, 3 \\ 0 & \text{else.} \end{cases}$$

and

$$p_{X|Y}(x|1) = \begin{cases} \frac{2}{3} & x = 2 \\ \frac{1}{3} & x = 3 \\ 0 & \text{else.} \end{cases}$$

In looking at these conditional probabilities, we see that $p_{X|Y}(x|1)$ is markedly different from $p_{X|Y}(x|0)$ This is partially due to the fact that, if we know that $Y = 1$ (the roll of the dice is even) it isn't possible that $X = 1$. That is, certain values of X are inaccessible depending on what we know about Y.

In looking at the previous example, we observe that the conditional probability mass function is always non-negative and

$$\sum_x p_{X|Y}(x|y) = 1$$

for each value of y with $p_Y(y) > 0$. This observation is really nothing more than we had previously established in Proposition 4.2. Its interpretation is that, given any non-empty event $Y = y$, $x \mapsto p_{X|Y}(x|y)$ is a probability mass function itself. It describes what we can know about the values of X given that we know Y takes the value Y.

Proposition 6.23. *Let X and Y be random variables on a common sample space Ω which is equipped with probability measure $\mathbb{P}$.*

1. *If X and Y are discrete and $p_Y(y) = \mathbb{P}(Y = y) > 0$, then $p(x) = p_{X|Y}(x|y)$ is a probability mass function.*
2. *If X and Y are jointly continuous and $f_Y(y) > 0$, then $f(x) = f_{X|Y}(x|y)$ is a probability density function.*

Should put laws of total probability here.

We recall that, if A and B are independent events with $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. In other words, we learn nothing about A from knowing B when A and B are dependent. In terms of random variables, we have the following result as an immediately consequence of Propositions 6.14 and 6.15.

Proposition 6.24. *Let X and Y be independent random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$.*

1. *If X and Y are discrete, then, for each y such that $p_Y(y) = \mathbb{P}(Y = y) > 0$,*

$$p_{X|Y}(x|y) = p_X(x)$$

for all x.

2. *If X and Y are jointly continuous, then, for each y such that $f_Y(y) > 0$,*

$$f_{X|Y}(x|y) = f_X(x)$$

for all x.

Example 6.16: Two Independent Coin Flips

Consider the random variables X_1 and X_2 on the sample space of two independent and fair coin flips presented in Example 6.2 (and revisited in Examples 6.7 and 6.14). We recall that $X_k = 1$ if the kth flip was heads and 0 if it was tails for $k = 1, 2$. There, we had

$$p_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{4} & (x_1, x_2) = (0,0), (1,0), (0,1), (1,1) \\ 0 & \text{else} \end{cases}$$

and X_1 and X_2 were Bernoilli random variables with parameter $1/2$, i.e., $p_{X_1}(k) = p_{X_2}(k) = 1/2$ when

$k = 0, 1$ and 0 otherwise. With this, we see that, for $x_2 = 0, 1$,

$$p_{X_1|X_2}(x_1|x_2) = \frac{p_{X_1,X_2}(x_1,x_2)}{p_{X_2}(x_1)} = \begin{cases} \frac{1/4}{1/2} & x_1 = 0,1 \\ 0 & \text{else} \end{cases} = p_{X_1}(x_1).$$

In view of the preceding proposition, this is consistent with our conclusion that $X_1 \perp X_2$ which we found in Example 6.7.

Example 6.17: I

n Example 6.3, we modeled the landing position of marbles in a box by jointly continuous random variables X and Y with

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{24} & -2 \leq x \leq 2,\ -3 \leq y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

for $(x, y) \in \mathbb{R}^2$. In Example 6.5, we found that $X \sim \text{Unif}([-2, 2])$ and $Y \sim \text{Unif}([-3, 3])$. For $-3 \leq y \leq 3$ with $f_Y(y) = 1/6 > 0$, observe that

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1/24}{1/6} = \frac{1}{4}$$

for $-2 \leq x \leq 2$ and 0 otherwise. Hence, $f_{X|Y}(x, y) = f_X(x)$ for each $-3 \leq y \leq 3$ This confirms our suspicion that the x and y-landing positions of the marbles are independent. Note here.

Example 6.18: Dart-Throwing Models

Let's consider our dart-throwing model described in Example 6.6 which modeled the landing position of a dart by (X, Y) for a pair of jointly continuous random variables X and Y. We had joint probability density function

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \sqrt{x^2+y^2} \leq 1 \\ 0 & \text{else} \end{cases}.$$

and marginal densities

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} & -1 \leq x \leq 1 \\ 0 & \text{else} \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} \frac{2}{\pi}\sqrt{1-y^2} & -1 \leq y \leq 1 \\ 0 & \text{else} \end{cases}$$

for $x, y \in \mathbb{R}$. For fixed $-1 < y < 1$, we have $f_Y(y) > 0$ and so

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{2\sqrt{1-y^2}} & -\sqrt{1-y^2} \leq x \leq \sqrt{1-y^2} \\ 0 & \text{else} \end{cases}.$$

Let's make two observations about this conditional density function. First, it is clear that $f_{X|Y}(x|y) \neq f_X(x)$ and, with this and in view of the proposition above, we have another confirmation that X and Y are not independent. Second, for fixed $-1 < y < 1$, $f(x) = f_{X|Y}(x|y)$ is the density of a continuous uniform random variables on the interval $[-\sqrt{1-y^2}, \sqrt{1-y^2}]$. This should come at no surprise because our original joint density was uniform on the dart board $\mathcal{D}$ and, if we condition on the event that Y belongs to a section/sliver $Y = y$, then we do expect the random variables X (conditioned on $Y = y$) to be uniformly distributed to that section of the board.

Exercise 6.14: Conditioning of the Better-dart model

Consider the jointly continuous random variables X and Y studied in Exercise 6.4 with

$$f_{X,Y}(x,y) = \frac{8}{\pi}e^{-8(x^2+y^2)}$$

for $(x,y) \in \mathbb{R}^2$.

1. Using Proposition 6.15, show that X and Y are independent.
2. Confirm the assertion of the preceding proposition, i.e., that $f_{X|Y}(x|y) = f_X(x)$.

As an application of the preceding results, consider two independent discrete random variables X and Y and ask: What can we say about their sum $Z = X + Y$. It is clear that the random variables Z is discrete and we can use the law of total probability to understand its probability mass function. For $z \in \mathbb{R}$, we have

$$p_Z(z) = P(X+Y=z) = \sum_{y \in R(Y)} \mathbb{P}(X+Y=z|Y=y)\mathbb{P}(Y=y)$$

where we have partitioned the sample space by the events $\{Y=y\}$ for $y \in R(Y)$. Upon noting that

$$\mathbb{P}(X+Y=z|Y=y) = \mathbb{P}(X+y=z|Y=y) = \mathbb{P}(X=z-y|Y=y) = p_{X|Y}(z-y|y)$$

for each $y \in R(Y)$, we have

$$p_Z(z) = \sum_{y \in R(Y)} p_{X|Y}(z-y|y)p_Y(y).$$

In the case that X and Y are continuous, $p_{X|Y} = p_X$ and so

$$p_Z(z) = \sum_{y \in R(Y)} p_X(z-y)p_Y(y).$$

This summation is called the **convolution product of p_X and p_Y** and written $p_X * p_Y$. In other words, when X and Y are independent discrete random variables, then

$$p_{X+Y} = p_X * p_Y$$

where

$$(p_X * p_Y)(k) = \sum_y p_X(k-y)p_Y(y).$$

Example 6.19: Sums of Independent Bernoulli Random Variables

Let's apply what we learned above to independent Bernoulli random variables $X, Y \sim \text{Ber}(p)$. In this case, we have

$$p_X(k) = p_Y(k) = \begin{cases} p & k=1 \\ q=1-p & k=0 \\ 0 & \text{else} \end{cases}.$$

Consequently,

$$p_{X+Y}(k) = \sum_{y=0,1} p_X(k-y)p_Y(y) = p_X(k-0)p_Y(0) + p_X(k-1)p_Y(1) = p_X(k)q + p_X(k-1)p.$$

Evaluating this sum for various values of k (you should do this!), we find that it is only non-zero for $k = 0, 1, 2$

(which is precisely what we should expect as the range of $X+Y$) and

$$p_{X+Y}(k) = \begin{cases} q^2 & k=0 \\ 2pq & k=1 \\ p^2 & k=2 \\ 0 & \text{else} \end{cases} = \begin{cases} \binom{2}{k} p^k q^{2-k} & k=0,1,2 \\ 0 & \text{else.} \end{cases}$$

In other words, the sum is a binomial random variable $X+Y \sim \text{Bin}(2,p)$. Continuing this argument inductively (and you should try it), we find:

Proposition 6.25. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed Bernoulli random variables all with parameter p. Then*

$$S_n = X_1 + X_2 + \cdots + X_n$$

is a Binomial random variables with parameters n and p, i.e., $S_n \sim Bin(n,p)$.

The following proposition summarizes our argument before the example and puts it together with an analogous result for jointly continuous independent random variables.

Proposition 6.26. *Let X and Y be independent random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$.*

1. *If X and Y are discrete with marginals p_X and p_Y, then $X+Y$ is a discrete random variable with probability mass function*

$$p_{X+Y}(z) = (p_X * p_Y)(z) = \sum_y p_X(z-y)p_Y(y)$$

for $z \in \mathbb{R}$.

2. *If X and Y are jointly continuous with marginal densities f_X and f_Y, respectively, then $X+Y$ is a continuous random variable with density*

$$f_{X+Y}(z) = (f_X * f_Y)(z) = \int_{\mathbb{R}} f_X(z-y)f_Y(y)\,dy$$

Note here

Exercise 6.15: The sum of independent Poisson random variables is Poisson

Let X and Y be Poisson random variables with parameters λ_1 and λ_2. Show that, if X and Y are independent, $X+Y$ is also a Poisson random variable with parameter $\lambda_1+\lambda_2$. Hint: You will need to make use of the binomial theorem.

Exercise 6.16: The sum of independent normal random variables are normal

By using the convolution product, one can prove the following important result:

Theorem 6.27. *If X_1 and X_2 are independent normal random variables with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then X_1+X_2 is a normal random variable with mean $\mu = \mu_1+\mu_2$ and variance $\sigma^2 = \sigma_1^2+\sigma_2^2$.*

In this exercise, you will show this in the special case that X and Y are both standard normal random variables. To this end, let X and Y be independent random variables with $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. Show that $X+Y \sim \mathcal{N}(0,2)$ by the following steps:

1. Write down that convolution equation

$$f_{X+Y}(z) = \int_{\mathbb{R}} f_X(z-y) f_Y(y)\, dy$$

where f_X and f_Y are the pdfs corresponding to X and Y which are, by assumption, both standard normal random variables. Simplify the integrand as much as possible. This means combining exponents, simplifying and factoring out any terms that do not depend on the integration variable y.

2. Your simplified integrant should have exponent $-(y^2 - zy) = zy - y^2$. By completing the square, show that this can be written as

$$-(y - a_z)^2 + z^2/4.$$

Do this and identify a_z.

3. Substituting your new expression for $-y^2 + zy$, you can now factor out the term $e^{z^2/4}$ outside of the integral. You will be left a prefactor (a multiplicative factor depending on z) and the integral

$$\int_{\mathbb{R}} e^{-(y-a_z)^2}\, dy.$$

By making the change of variables $y \mapsto y' = y - a_z$, find the value of this integral.

4. Simplify your result to obtain $f_{X+Y}(z)$. Can you conclude that $X + Y$ is $\mathcal{N}(0, 2)$?

6.4.1 Conditional Expectation and Variance

Now that we have a good understanding of conditioning of random variables, at least in the discrete and continuous cases, we are in a position to discuss conditional expectation and variance. The former captures the essence of how we can update a prediction of X if we know something about Y.

Definition 6.28. *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$.*

1. *If X and Y are discrete, then the conditional expectation of X given that $Y = y$ is given by*

$$\mathbb{E}(X|Y = y) = \sum_x x \cdot p_{X|Y}(x|y)$$

for each y such that $p_Y(y) > 0$.

2. *If X and Y are jointly continuous, then the conditional expectation of X given that $Y = y$ is given by*

$$\mathbb{E}(X|Y = y) = \int_{\mathbb{R}} x f_{X|Y}(x|y)\, dx$$

whenever $f_Y(y) > 0$.

Example 6.20:

We return to our random variables X and Y of Example 6.1. In Example 6.15, we found that

$$p_{X|Y}(x|0) = \begin{cases} \frac{1}{3} & x = 1, 2, 3 \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad p_{X|Y}(x|1) = \begin{cases} \frac{2}{3} & x = 2 \\ \frac{1}{3} & x = 3 \\ 0 & x \text{ else} \end{cases}.$$

In view of the above definition, we have

$$\mathbb{E}(X|Y=0) = \sum_x x p_{X|Y}(x|0) = 1\left(\frac{1}{3}\right) + 2\left(\frac{1}{3}\right) + 3\left(\frac{1}{3}\right) = 2$$

and

$$\mathbb{E}(X|Y=1) = \sum_x x p_{X|Y}(x|1) = 2\left(\frac{2}{3}\right) + 3\left(\frac{1}{3}\right) = \frac{7}{3}.$$

Let's discuss why this should made some intuitive sense. If we condition on the event that $Y = 0$, we haven't really learned much about the value of X because $X = 1, 2$ and 3 are all equally likely and so our "best guess" at X should be the average of these numbers: 2. If we condition on the event $Y = 1$, we do know that $X = 1$ isn't possible and that $X = 2$ is twice as likely as $X = 3$. Hence our "best guess" should be somewhere between 2 and 3 but closer to 2 than 3 – this is consistent with our result of $\mathbb{E}(X|Y = 1) = 7/3 = 2 + 1/3$.

Exercise 6.17: Chemical Impurities

A company producing a chemical finds that there are two types of impurities (called Type 1 and Type 2) found in their product. For a certain volume they chemical, they find that they can model the proportion of total impurities by X and the proportion of Type 1 impurities by Y which are jointly continuous random variables with

$$f_{X,Y}(x,y) = \begin{cases} 2(1-x) & 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{else} \end{cases}.$$

1. Compute the marginal densities and expectations of the random variables X and Y.
2. For each y with $f_Y(y) > 0$, find the conditional expectation $\mathbb{E}(X|Y = y)$. Explain why this should give the "best guess" for X given the $Y = y$.

Now that we have a good understanding of the conditional expectation of a random variables X given the event $Y = y$, we are in a position to introduce the conditional expectation of X given Y as a random variable itself.

Definition 6.29. *Let X and Y be random variables on a common sample space Ω equipped with probability P. The conditional expectation of X given Y is the random variable $\mathbb{E}(X|Y)$ defined by*

$$\mathbb{E}(X|Y)(\omega) = \mathbb{E}(X|Y = Y(\omega))$$

for $\omega \in \Omega$. In other words, $\mathbb{E}(X|Y)$ is the random variable that takes the value $\mathbb{E}(X|Y = y)$ when $y = Y(\omega)$.

To get a better understanding of the random variable $\mathbb{E}(X|Y)$, let's take a look at two cases for random variables X and Y which we assume to be discrete. First, let's suppose that X and Y are independent random variables. In this case, $p_{X|Y}(x|y) = p_X(y)$ for each $y \in R(Y)$ and so

$$\mathbb{E}(X|Y = y) = \sum_x x p_X(x) = \mathbb{E}(X)$$

for each $y \in R(Y)$. Since this does not depend on y at all, the conditional expectation of X given Y is simply the constant random variable: $\mathbb{E}(X|Y)(\omega) = \mathbb{E}(X)$. Our interpretation here is that, if X and Y are independent, the "best guess" for X given Y is simply the expected value of X. Since we can't really learn anything about X from knowing something about Y, we can't really do any better.

At the other extreme from independence is the case in which X is a function of Y, i.e., $X = \varphi(Y)$. In this case

$$\begin{aligned} p_{X|Y}(x|y) &= \mathbb{P}(X = x|Y = y) \\ &= \mathbb{P}(\varphi(Y) = x|Y = y) \\ &= \mathbb{P}(\varphi(y) = x|Y = y) \\ &= \begin{cases} 1 & \varphi(y) = x \\ 0 & \text{otherwise} . \end{cases} \end{aligned}$$

From this it follows that, for each $y \in R(Y)$,

$$\mathbb{E}(X|Y = y) = \sum_x p_{X|Y}(x|y) = \varphi(y)$$

since the only non-zero term in this sum is that for which $x = \varphi(y)$ and $p_{X|Y}(x|y) = p_{X|Y}(\varphi(y)|y) = 1$. Consequently, for each $\omega \in \Omega$,

$$\mathbb{E}(X|Y)(\omega) = \mathbb{E}(X|Y = Y(\omega) = \varphi(Y(\omega)) = X(\omega).$$

In other words, if X is a function of Y, then the "best guess" for X given Y is just X itself.

A key property of conditional expectation is the following result which says that, if we average the conditional expectation $\mathbb{E}(X|Y)$ over all possibilities of ω and hence values of Y, we simply recover the expectation of X.

Theorem 6.30 (The Law of Iterated Expectation)**.** *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Then the random variable $\mathbb{E}(X|Y)$ satisfies,*

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

Proof. Here we give a proof for the case in which X and Y are discrete. Since $\mathbb{E}(X|Y)$ can be viewed as a function of Y, i.e., $\varphi(Y) = \mathbb{E}(X|Y)$ where $\varphi(y) = \mathbb{E}(X|Y = y)$, we have

$$\mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \mathbb{E}(X|Y = y)p_Y(y)$$

by virtue of Theorem 5.41. Using the definition of conditional expectation of X given $Y = y$ and the definition of $p_{X|Y}$, we have

$$\begin{aligned} \mathbb{E}(\mathbb{E}(X|Y)) &= \sum_y \sum_x x p_{X|Y}(x|y)\, p_Y(y) \\ &= \sum_{x,y} x\, p_{X,Y}(x, y) \\ &= \sum_x x \sum_y p_{X,Y}(x, y) \\ &= \sum_x x\, p_X(x) \\ &= \mathbb{E}(X) \end{aligned}$$

as was asserted. □

Example 6.21:

Let's consider the random variables X and Y of Example 6.1 in which we computed

$$p_X(x) = \begin{cases} \frac{1}{6} & x = 1 \\ \frac{1}{2} & x = 2 \\ \frac{1}{3} & x = 3 \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad p_Y(y) = \begin{cases} \frac{1}{2} & y = 0, 1 \\ 0 & \text{else} \end{cases}.$$

In Example 6.20, we found that

$$\mathbb{E}(X|Y = y) = \begin{cases} 2 & y = 0 \\ \frac{7}{3} & y = 1 \end{cases}$$

and therefore

$$\mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \mathbb{E}(X|Y = y)p_Y(y) = 2\left(\frac{1}{2}\right) + \frac{7}{3}\left(\frac{1}{2}\right) = \frac{13}{6}.$$

Also, we have

$$\mathbb{E}(X) = \sum_x p_X(x) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{2}\right) + 3\left(\frac{1}{3}\right) = \frac{1}{6} + 2 = \frac{13}{6}$$

as it must be in view of the law of iterated expectation.

Exercise 6.18: More Chemical Impurities

Confirm the law of iterated expectation for the random variables X and Y in Exercise 6.17.

To accompany the law of iterated expectation, we have an analogous theorem for conditional variance.

Theorem 6.31. *Let X and Y be random variables on a common sample space Ω equipped with probability measure $\mathbb{P}$. Then the conditional variance of X given Y is the random variable defined by*

$$\text{Var}(X|Y) = \mathbb{E}((X - \mathbb{E}(X|Y)^2|Y)$$

and satisfies

$$\text{Var}(X) = \mathbb{E}(\text{Var}(X|Y)) - \text{Var}(\mathbb{E}(X|Y)).$$

Chapter 7

Limit Theorems: The law of large numbers and the central limit theorem

Our goal in this chapter is to understand what happens when we sum large numbers of independent and identically distributed random variables. To be precise, let $X_1, X_2, X_3, \dots$ be a sequence of independent and identically distributed random variables (henceforth abbreviated **i.i.d**) on a common sample space Ω equipped with a probability measure $\mathbb{P}$. Together, Ω and $\mathbb{P}$ can be thought of as modeling an experiment involving (an infinite number of) independent trials. In this case, the random variables $X_1, X_2, \dots,$ being identically distributed, can be though of as encoding exactly the same information but where each X_k is only sensitive to what happens on the kth trial. For each $n \in \mathbb{N}_+ = \{1, 2, \dots, \}$, define

$$S_n = X_1 + X_2 + \cdots + X_n$$

and

$$\overline{X_n} = \frac{1}{n} S_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

to be the sum and corresponding so-called **statistical average** of the first n random variables. In these terms, our goal in this chapter is to say something about S_n and $\overline{X_n}$ as $n \to \infty$. Let us motivate this study by considering two examples with which we are already very familiar.

Example 7.1: Alice's Random Walk

Once again, let's consider Alice's walk along the integers $\mathbb{Z}$. Generalizing the ideas of Exercise 5.6, we suppose that Alice starts her journey at $x = 0$ and flips a biased coin with probability p of "heads" and $q = 1 - p$ of "tails" where $0 \leq p \leq 1$. We assume that she flips this coin over and over again independently and, after each flip, moves to the right or left according to whether "heads" or "tails" appears. This random walk can be modeled by considering the sample space

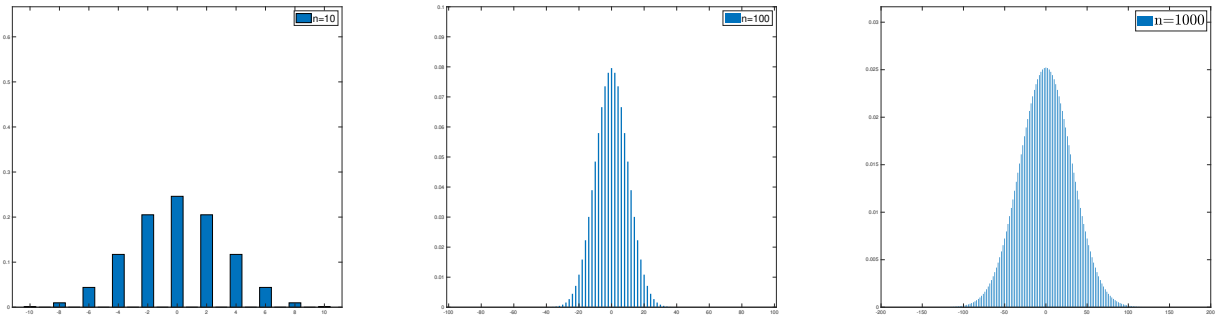
$$\Omega = \{\omega = (\omega_1, \omega_2, \omega_3, \dots) : \omega_k = \text{ "heads" or "tails" for each } k = 1, 2, \dots\}$$

and the random variables $X_1, X_2, \dots$ defined, for each $k \in \mathbb{N}_+$, by

$$X_k(\omega) = \begin{cases} 1 & \omega_k = \text{"heads"} \\ -1 & \omega_k = \text{"tails"} \end{cases}$$

for $\omega = (\omega_1, \omega_2, \dots) \in \Omega$; we note that each random variable X_k only depends on ω_k, i.e., it only depends on the result of the kth coin flip (or trial). To describe the appropriate probability measure $\mathbb{P}$ modeling repeated independent flips of the same coin, it is enough to ask that

$$\mathbb{P}(X_k = 1) = \mathbb{P}(\{\omega \in \Omega : \omega_k = \text{ "heads"}\}) = p,$$



(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$

Figure 7.1: $\mathbb{P}(S_n = x)$ for $n = 10, 100$, and 1000.

$$\mathbb{P}(X_k = -1) = \mathbb{P}(\{\omega \in \Omega : \omega_k = \text{“tails”}\}) = q,$$

for each $k \in \mathbb{N}_+$ and that the random variables $X_1, X_2, \ldots,$ (and hence the trials) are independent. Intuitively, this probability measure $\mathbb{P}$ is consistent with the assumption that Alice's coin flips are done identically and independently; however, we note that, along the lines of our discussion at the end of Section 6.2, the fact that such a probability measure exists (and the above is exactly what's needed to specify it uniquely) is a highly non-trivial matter.

With this collection of i.i.d random variables, we observe that

$$S_n = X_1 + X_2 + \cdots X_n$$

is precisely Alice's position after n flips of the coin. To understand Alice's walk, we are interested in the following questions:

- What happens to the statistical average $\overline{X_n}$ of Alice's steps as $n \to \infty$?
- How can we understand Alice's position S_n after a large number of steps? For example, given $x \in \mathbb{Z}$, what is the probability that Alice is at position x after n steps? Equivalently, what is $\mathbb{P}(S_n = x)$?

As we will see, these questions are best answered using the major theorems of this chapter, the law of large numbers and the central limit theorem. To have an idea of what we can expect of the second question, Figure 7.1 illustrates $\mathbb{P}(S_n = x)$ for several values of n in the case that $p = q = 1/2$. Though S_n is discrete, notice how similar the plots start to look like the density of a normal random variable.

Exercise 7.1:

For the random variables $S_n = X_1 + X_2 + \cdots + X_n$ and $\overline{X_n} = S_n/n$ in Alice's random walk, show that, for each n, $\mathbb{E}(S_n) = n(p - q)$, $\mathbb{E}(\overline{X_n}) = p - q$ and $\text{Var}(S_n) = 4npq$.

Example 7.2: Frequentist Interpretation

Let's consider an infinite collection of independent trials of an experiment. For example, we could roll a perfect die over and over again independently ad infinitum. Our goal in this example is to understand how we can count the number of times a certain event occurs while performing the same experiment over and over again independently.

Mathematically, our infinite collection of trials can be represented by considering a "base" sample space Ω which is equipped with a probability measure $\mathbb{P}$. To model an infinite number of trials, we form the sample space

$$\Omega^\infty = \{\omega = (\omega_1, \omega_2, \omega_3, \dots) : \omega_k \in \Omega \text{ for each } k = 1, 2, \dots\}$$

where each $\omega = (\omega_1, \omega_2, \dots)$ represents a single outcome of performing the experiment over and over again. Though it is a deep mathematical result, it is a fact that there is a probability measure $\mathbb{P}_\infty$ on Ω^∞ which assigns the probability

$$\mathbb{P}_\infty(R) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdots \mathbb{P}(A_n)$$

to every event in Ω^∞ of the form

$$R = \{\omega = (\omega_1, \omega_2, \dots) \in \Omega : \omega_k \in A_k \text{ for each } k = 1, 2, \dots, n\}.$$

The event R is called a **finite rectangle** and consists precisely of those outcomes in Ω^∞ in which, for each $k = 1, 2, \dots, n$, the subevent A_k occurs in the kth trial (and allows for anything to happen beyond the nth trial). This probability measure $\mathbb{P}_\infty$ is exactly what makes the trials independent and it is called an infinite product measure.

For example, in our die-rolling situation, if we take the base sample space to be $\Omega = \{1, 2, 3, 4, 5, 6\}$ equipped with the uniform measure $\mathbb{P}$, then Ω^∞ represents an infinite sequence of rolls of the die. Consider the finite rectangular event

$$R = \{\omega = (\omega_1, \omega_2, \dots) \in \Omega^\infty : \omega_1 \in \{1, 2\}, \omega_2 \in \{3, 4\}, \text{ and } \omega_3 \in \{5\}\};$$

this is the event that we get 1 or 2 on the first roll, 3 or 4 on the second roll, exactly 5 on the third roll, and then we stop paying attention. In this case,

$$\mathbb{P}_\infty(R) = \mathbb{P}(\{1, 2\})\mathbb{P}(\{3, 4\})\mathbb{P}(\{5\}) = \left(\frac{2}{6}\right)\left(\frac{2}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{54}.$$

Back to the general setting, consider an event $A \subseteq \Omega$ and consider the sequence of random variables $X_1, X_2, \dots$, on Ω^∞ defined, for each $n = 1, 2, \dots$, by

$$X_n(\omega) = \mathbb{1}_A(\omega_n) = \begin{cases} 1 & \omega_n \in A \\ 0 & \omega_n \notin A \end{cases}$$

for $\omega = (\omega_1, \omega_2, \dots) \in \Omega^\infty$. Each random variable X_n simply indicates if A occurs on the nth trial. By construction, the random variables (each of which just depends on the outcome of a single trial) are all independent (You should try to verify this if you're interested – you have all the tools!). Also, observe that each random variable X_n is discrete taking only the values 0 and 1. I claim that $X_1, X_2, \dots$ are all Bernoulli random variables with the same parameter $p = \mathbb{P}(A)$ and so, in particular, they are identically distributed. To see this, first observe that, for each n, the event $\{X_n = 1\}$ is a finite rectangle because it can be written in the form

$$\{X_n = 1\} = \{\omega_n \in A\} = \{\omega \in \Omega^\infty : \omega_n \in A \text{ and } \omega_j \in \Omega \text{ for all } j \neq n\}.$$

Therefore

$$\mathbb{P}_\infty(\{X_n = 1\}) = \mathbb{P}(\Omega) \cdot \mathbb{P}(\Omega) \cdots P(\Omega) \cdot \mathbb{P}(A) = 1 \cdot 1 \cdots 1 \cdot \mathbb{P}(A) = \mathbb{P}(A)$$

and so $\mathbb{P}_\infty(X_n = 0) = 1 - \mathbb{P}(A)$. Thus, for each n, $X_n \sim \text{Ber}(p)$ where $p = \mathbb{P}(A)$ which proves our claim. Consequently, the collection $X_1, X_2, \dots$ is a collection of independent and identically distributed Bernoilli random variables with common mean $\mu = p = \mathbb{P}(A)$.

Let's now consider the statistical mean

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

which assigns, to each $\omega \in \Omega^\infty$ the value

$$\overline{X_n}(\omega) = \frac{\mathbb{1}_A(\omega_1) + \mathbb{1}_A(\omega_2) + \cdots \mathbb{1}_A(\omega_n)}{n}.$$

Upon noting that the terms in the numerator are either 1 or 0 according to whether or not A occurs on the kth trial for $k = 1, 2, \ldots, n$, we see that

$$\overline{X_n}(\omega) = \frac{\#(\text{Occurences of } A \text{ in first } n \text{ trials})}{n}.$$

In other words, $\overline{X_n} = S_n/n$ where $S_n \sim \text{Bin}(n, p)$ for $p = \mathbb{P}(A)$. To investigate the frequentist interpretation of probability discussed in Chapter 2, we would like to understand the behavior of this statistical average $\overline{X_n}$ as $n \to \infty$.

With the above examples in mind, our goal is to understand the large n behavior of S_n and $\overline{X_n}$. In order to do this in a meaningful way, we must first discuss convergence.

7.0.1 Types of convergence

In contrast to sequences of numbers, there are many different ways (in fact, an infinite number of ways) in which a sequence of random variables can converge to another random variable. To this end, we give a definitions that presents three inequivalent notions of convergence, the first is the strongest and is, in fact, rarely satisfied.

Definition 7.1 (Three notions of convergence)**.** *Let $Y_1, Y_2, \ldots$ be a sequence of random variables defined on a common sample space Ω equipped with probability measure $\mathbb{P}$. We shall denote by $F_n = F_{Y_n}$ the cumulative distribution function of Y_n for each $n \in \mathbb{N}$. Also, let Y be another random variable on Ω and denote by $F = F_Y$ its cumulative distribution function.*

1. *We say that* ***the sequence*** Y_n ***converges to*** Y ***almost surely*** *if the event*

$$\left\{\lim_{n\to\infty} Y_n = Y\right\} = \left\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\}$$

has probability 1*, i.e.,*

$$\mathbb{P}\left(\lim_{n\to\infty} Y_n = Y\right) = 1.$$

In this case, we shall write $Y_n \overset{a.s.}{\Longrightarrow} Y$.

2. *We say that* ***the sequence*** Y_n ***converges to*** Y ***in probability*** *if, for each $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| < \epsilon) = 1 \qquad \textit{or, equivalently,} \qquad \lim_{n\to\infty} \mathbb{P}(|Y_n - Y| \geq \epsilon) = 0.$$

In this case, we shall write $Y_n \overset{\mathbb{P}}{\Longrightarrow} Y$.

3. *We say that* ***the sequence*** Y_n ***converges to*** Y ***in distribution*** *if*

$$\lim_{n\to\infty} F_n(y) = F(y)$$

for each point y at which F is continuous. In this case, we shall write $Y_n \overset{d}{\Longrightarrow} Y$. Convergence in distribution is also referred to as weak convergence, the use of vocabulary being due to an equivalent measure-theoretic notion.

Taken all at once, the above definition can be intimidating. Though each notion does give a very precise way in which a sequence of random variables converges, the main idea is that each captures a way in which, asymptotically, the random variables/functions Y_n are close to Y. Let's consider three examples.

Example 7.3: A simple example of convergence

Consider the sample space $\Omega = \{H, T\}$ and the probability measure $\mathbb{P}$ assigning $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1 - p$ and modeling a single biased coin flip; here, we assume $0 < p < 1$. On Ω, we can consider a sequence of random variables Y_n defined, for each $n \in \mathbb{N}_+$, by

$$Y_n(\omega) = \begin{cases} 1 + \frac{1}{n} & \omega = H \\ 0 & \omega = T \end{cases}.$$

I claim that Y_n converges to the Bernoulli random variable

$$Y(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$$

in distribution, in probability, and almost surely.

Let's start with convergence in distribution. We can easily compute

$$F_n(y) = F_{Y_n}(y) = \mathbb{P}(Y_n \leq y) = \begin{cases} 0 & y < 0 \\ q & 0 \leq y < 1 + \frac{1}{n} \\ 1 & y \geq 1 + \frac{1}{n} \end{cases}$$

and

$$F(y) = F_Y(y) = \begin{cases} 0 & y < 0 \\ q & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

for $y \in \mathbb{R}$. In looking at $F(y)$, we see that it has discontinuities at $y = 0$ and $y = 1$. For $y < 0$,

$$\lim_{n\to\infty} F_n(y) = \lim_{n\to\infty} 0 = 0 = F(y).$$

For $0 < y < 1$, we have $F_n(y) = q$ for all n since $y < 1 + 1/n$ and so

$$\lim_{n\to\infty} F_n(y) = \lim_{n\to\infty} q = q = F(y).$$

For $y > 1$, we note that $1 + 1/n \leq y$ for sufficiently large n and therefore

$$\lim_{n\to\infty} F_n(y) = \lim_{n\to\infty} 1 = 1 = F(y)$$

Thus, $\lim_{n\to\infty} F_n(y) = F(y)$ at all points where $F(y)$ is continuous and therefore Y_n converges to Y in distribution. It should be noted that $F_n(1) = q \neq 1 = F(1)$ and so, in particular, $\lim_{n\to\infty} F_n(1) \neq F(1)$. This, of course, doesn't rule out convergence in distribution because 1 is not a point of continuity of F. As Y_n really does converge to Y in all of the senses claimed, this illustrates why we really ask only for the limit at points of continuity – we wouldn't want to rule out such a triviality.

Let's now discuss convergence in probability. Let $\epsilon > 0$ and observe that

$$|Y_n(T) - Y(T)| = |0 - 0| = 0 < \epsilon$$

for all n. When $n > 1/\epsilon$,

$$|Y_n(H) - Y(H)| = \left|1 + \frac{1}{n} - 1\right| = \frac{1}{n} < \epsilon$$

Thus, whenever $n > 1/\epsilon$,

$$\{\omega = \Omega : |Y_n(\omega) - Y(\omega)| < \epsilon\} = \Omega$$

and hence

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| < \epsilon) = \mathbb{P}(\Omega) = 1.$$

Thus, Y_n converges to Y in probability.

Finally, we discuss convergence almost surely. When $\omega = H$, we have

$$\lim_{n\to\infty} Y_n(\omega) = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right) = 1 + 0 = Y(\omega)$$

and, for $\omega = T$,

$$\lim_{n\to\infty} Y_n(\omega) = \lim_{n\to\infty} 0 = 0 = Y(\omega).$$

Thus,

$$\left\{\lim_{n\to\infty} Y_n = Y\right\} = \Omega$$

and so this event must have probability 1. In other words, we have shown that Y_n converges to Y almost surely.

Example 7.4: The maximum of independent uniform random variables

Let $X_1, X_2, \ldots,$ be a sequence of independent random variables all distributed uniformly on the interval $[0, 1]$. For each n, define

$$Y_n = \max\{X_1, X_2, \ldots, X_n\}.$$

To understand the random variable Y_n, observe that, for ω, $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$ is a list of real numbers in the interval $[0, 1]$. Since these random variables all are uniformly distributed in the interval $[0, 1]$, each has a non-zero probability of being in any interval of the form $[1 - \epsilon, 1]$ for any small number $\epsilon > 0$. Since this sequence is independent, it is reasonable to suspect that, by consider a large number n of them, there is a high probability of having at least one of them take its value close to 1. Thus, we expect that their maximum, Y_n should be close to 1 with high probability as long as n is large. In thinking about this a little big more, we suspect that the random variables, at least on average, get close to the constant random variable

$$Y = 1.$$

In other words, we suspect that the sequence of random variables Y_n converge to $Y = 1$ in some sense. Let's investigate this conjecture in the sense of convergence in distribution.

For $y \in \mathbb{R}$,

$$F_n(y) = F_{Y_n}(y) = \mathbb{P}(Y_n \leq y) = \mathbb{P}(X_1 \leq y, X_2 \leq y, \ldots, X_n \leq y)$$

since $Y_n \leq y$ if and only if $X_k \leq y$ for all $k = 1, 2, \ldots, n$. Using the independence of the random variables $X_1, X_2, \ldots,$ we have

$$F_n(y) = \mathbb{P}(X_1 \leq y)\mathbb{P}(X_2 \leq y)\cdots\mathbb{P}(X_n \leq y) = F_X(y)^n$$

where

$$F_X(y) = \begin{cases} 0 & y < 0 \\ y & 0 \leq y \leq 1 \\ 1 & y > 1 \end{cases}$$

is the common cumulative distribution function for the random variables $X_1, X_2, \ldots$. Observe that, for $0 < y < 1$,

$$\lim_{n\to\infty} F_n(y) = \lim_{n\to\infty} F_X(y)^n = \lim_{n\to\infty} y^n = 0.$$

By performing a similar calculation on the intervals $(-\infty, 0]$ and $[1, \infty)$ we find that

$$\lim_{n\to\infty} F_n(y) = \begin{cases} 1 & y \geq 1 \\ 0 & y < 1 \end{cases}$$

for $y \in \mathbb{R}$. Let us note that the constant random variable $Y = 1$ has

$$F(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(1 \leq y) = \begin{cases} 1 & y \geq 1 \\ 0 & y < 1 \end{cases}$$

for $y \in \mathbb{R}$ and so

$$\lim_{n\to\infty} F_n(y) = F(y)$$

for each $y \in \mathbb{R}$. In particular, we conclude that Y_n converges to Y in distribution.

Let's now investigate convergence in probability. In light of our calculation above, we found that

$$F_n(y) = \begin{cases} 0 & y < 0 \\ y^n & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$$

for $y \in \mathbb{R}$. Now, given $\epsilon > 0$, consider the event

$$\{|Y_n - Y| \geq \epsilon\} = \{|Y_n - 1| \geq \epsilon\} = \{1 - Y_n \geq \epsilon\} = \{Y_n \leq 1 - \epsilon\}.$$

Using the cumulative distribution function of Y_n given above, we have

$$\mathbb{P}(|Y_n - Y| \geq \epsilon) = \mathbb{P}(Y_n \leq 1 - \epsilon) = \begin{cases} (1-\epsilon)^n & \epsilon \leq 1 \\ 0 & \epsilon > 1 \end{cases}$$

Since $\epsilon > 0$, it is clear that

$$\lim_{n\to\infty} (|Y_n - Y| \geq \epsilon) = 0$$

and from this we conclude that Y_n converges to Y in probability.

Finally, let's investigate almost sure convergence. To this end, I claim that

$$\{\lim_{n\to\infty} Y_n \neq Y\} \subseteq \bigcup_{n=1}^{\infty} \left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right); \tag{7.1}$$

in fact, these sets are equal but we'll only need to use the containment stated above. To see this, let $\omega \in \Omega$ be an outcome for which

$$\lim_{n\to\infty} Y_n(\omega) \neq Y(\omega) = 1.$$

Given that $Y_n(\omega)$ is formed by taking a maximum, $Y_n(\omega)$ is necessarily an increasing sequence and so the only way for the above limit to not be 1 is for

$$X_m(\omega) \leq Y_m(\omega) \leq 1 - \epsilon < 1$$

for all m where $\epsilon > 0$. In other words, it must be true that, for some n (with $1/n < \epsilon$),

$$X_m(\omega) \leq 1 - \frac{1}{n}$$

for all m and hence

$$\omega \in \bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \subseteq \bigcup_{n=1}^{\infty} \left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right)$$

which guarantees (7.1). With this in hand, observe that, for any $n, M \in \mathbb{N}_+$, the independence of the random variables $X_1, X_2, \ldots$ give

$$\begin{aligned} \mathbb{P}\left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right) &\leq \mathbb{P}\left(\bigcap_{m=1}^{M} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right) \\ &= \prod_{m=1}^{M} \mathbb{P}\left(X_m \leq 1 - \frac{1}{n} \right) \\ &= \prod_{m=1}^{M} \left(1 - \frac{1}{n} \right) \\ &= \left(1 - \frac{1}{n} \right)^M. \end{aligned}$$

Since $0 \leq 1 - 1/n < 1$, the only way for this to hold for all M is for

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right) = 0.$$

Thus, by monotonicity and the so-called union bound (Theorem 2.11), we have

$$\mathbb{P}(\lim_{n\to\infty} Y_n \neq Y) \leq \mathbb{P}\left(\bigcup_{n=1}^{\infty} \left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right) \right) \leq \sum_{n=1}^{\infty} \mathbb{P}\left(\bigcap_{m=1}^{\infty} \left\{ X_m \leq 1 - \frac{1}{n} \right\} \right) = \sum_{n=1}^{\infty} 0 = 0.$$

Thus,

$$\mathbb{P}(\lim_{n\to\infty} Y_n = Y) = 1 - \mathbb{P}(\lim_{n\to\infty} Y_n \neq Y) = 1$$

from which we conclude that Y_n converges to $Y = 1$ almost surely.

Example 7.5: The typewriter

Let $\Omega = [0, 1]$ which we take to be equipped with the probability measure $\mathbb{P}$ given, for each event $A \subseteq \Omega$, by $\mathbb{P}(A) = \int_A 1\, dx = \text{length}(A)$. For $n \in \mathbb{N}_+$, define

$$Y_n(\omega) = \mathbb{1}_{A_n}(\omega) = \begin{cases} 1 & \omega \in A_n \\ 0 & \text{else} \end{cases}$$

for $\omega \in \Omega$ where $A_1, A_2, \ldots$ is a sequence of events defined in the following (somewhat complicated) way: For $n \in \mathbb{N}$, set

$$A_n = \left[\frac{k}{2^m}, \frac{k+1}{2^m} \right]$$

where $m = \lfloor \log_2(n) \rfloor$ and $k = n - 2^m$. To get a feel for these events, observe that for $n = 1$, $m = \lfloor \log_2(1) \rfloor = 0$, $k = 1 - 2^0 = 0$ and so

$$A_1 = [0, 1] = \Omega.$$

For $n = 2$ and $n = 3$, we have

$$A_2 = [0, 1/2] \qquad \text{and} \qquad A_3 = [1/2, 1].$$

Continuing on, the events A_n are subintervals of $[0, 1]$ that move like a typewriter back and forth and at each level have smaller and smaller length; to get a feeling for them, I encourage you to write down A_4, A_5, A_6, A_7 and A_8 and graph the corresponding random variables Y_n for $n = 4, 5, 6, 7$ and 8. For any $0 < \epsilon < 1$,

$$\mathbb{P}\left(|Y_n - 0| \geq \epsilon\right) = \frac{1}{2^m} \leq \frac{2}{n}$$

where $m = \lfloor \log_2(n) \rfloor$ and we have used the fact that $\log_2(n) \leq m+1$ to see that $1/2^m \leq 2/n$. Consequently,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - 0| \geq \epsilon) = 0$$

and so Y_n converges to 0 in probability. However, observe that, for each $\omega \in [0, 1]$, there is an infinite subsequence of events A_n containing ω and another infinite subsequence of events that do not. Correspondingly, for each ω, $\lim_{n\to\infty} Y_n(\omega)$ does not exist and, in particular,

$$\mathbb{P}(\lim_{n\to\infty} Y_n = 0) = \mathbb{P}(\varnothing) = 0$$

showing that Y_n does not converge to 0 almost surely (it converges nowhere). This example shows that almost sure convergence and convergence in probability are not equivalent.

If you take a subsequent course in probability or a course in measure theory, you will spend a significant amount of time on these concepts (and other notions of convergence). For our purposes, it is enough to know the following result whose proof can be found in [1] or [3].

Proposition 7.2. *Convergence almost surely implies convergence in probability which, in turn, implies convergence in distribution. None of the reverse implications hold, i.e., there is a sequence of random variables which converges in distribution but does not converge in probability and, further, there is a sequence of random variables which converges in probability but not almost surely.*

7.1 The law of large numbers

Now that we have a precise way to discuss that manner in which random variables can converge, we are in a position to state our first main theorem. This theorem was originally established by Jacob Bernoulli in the simple case of independent random variables taking only the values 0 and 1 – we now call these Bernoulli random variables. Strengthened by subsequent work by Montmort, De Moivre, Stirling, Laplace, Poisson, Chebyshev, and Bienaymé [13], the following modern form of Bernoulli's theorem is called the weak law of large numbers[1].

Theorem 7.3 (Weak Law of Large Numbers)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. On Ω, let $X_1, X_2, X_3, \ldots$ be an i.i.d sequence of random variables with common mean μ and variance $\sigma^2 < \infty$. Then*

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

converges in probability to the constant random variable μ, i.e., $\overline{X_n} \overset{\mathbb{P}}{\Longrightarrow} \mu$.

The result above is remarkable. In its essence, it says that averaging a large number if i.i.d. random variables – no matter how they are distributed (so as long as they have a finite variance) – tends to a single (non-random) number $\mu = \mathbb{E}(X)$. It should be noted that the way that we defined $\mathbb{E}(X)$ has nothing, a priori, to do with averages.

[1] According to [13], it was Poisson who first used the term "law of large numbers". Bienaymé strongly dissaproved of this term.

It is simply given by the mathematical construction of the expectation based, essentially, on the probability measure $\mathbb{P}$ and, correspondingly, the distribution of X. Thus, the weak law of large numbers is our first real link to connecting the concept to expectation to that of the average; it is why we call $\mu = \mathbb{E}(X)$ the mean.

The finite-variance assumption in Theorem 7.3 can be weakened to $\mathbb{E}(|X|) < \infty$ and the result still holds; the proof, however, is considerably more difficult need to check this [3]. The proof given below, which assumes $\sigma^2 < \infty$, is essentially due the methods of Bienaymé and Chebyshev presented in the mid-nineteenth century [13]. We first treat two lemmas; the first is the so-called Markov inequality which you established in the context of a countable sample space in Exercise 5.11.

Lemma 7.4 (Markov's Inequality)**.** *Let X be a non-negative random variable on a sample space Ω with probability measure $\mathbb{P}$. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Lemma 7.5 (The Bienaymé-Chebyshev Inequality)**.** *Let X be a random variable on a sample space Ω equipped with probability measure $\mathbb{P}$. If X has mean μ and finite variance $\sigma^2 < \infty$, then, for every $\epsilon > 0$,*

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

Proof. Let $\epsilon > 0$ and observe that

$$\mathbb{P}(|X - \mu| \geq \epsilon) = \mathbb{P}((X - \mu)^2 \geq \epsilon^2) \leq \frac{\mathbb{E}((X - \mu)^2)}{\epsilon^2}$$

thanks to Markov's inequality and the fact that $(X - \mu)^2$ is a non-negative random variable. Since $\mathbb{E}((X - \mu)^2) = \text{Var}(X) = \sigma^2$, we have

$$\mathbb{P}(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2},$$

as desired. □

We are now in a position to prove the weak law of large numbers.

Proof of Theorem 7.3. By the linearity of expectations, we first observe that

$$\mathbb{E}(\overline{X}_n) = \frac{1}{n}\mathbb{E}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\left(\mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)\right) = \frac{1}{n}\left(\mu + \mu + \cdots + \mu\right) = \mu$$

where we have used our assumption that $X_1, X_2, \ldots$ are identically distributed with common mean μ. With the aim of applying the Bienaymé-Chebyshev inequality to the random variable $\overline{X_n}$, lets compute its variance. We have

$$\begin{aligned}
\text{Var}(\overline{X_n}) &= \mathbb{E}((\overline{X_n} - \mu)^2) \\
&= \mathbb{E}\left(\left(\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{n\mu}{n}\right)^2\right) \\
&= \frac{1}{n^2}\mathbb{E}\left((X_1 + X_2 + \cdots + X_n - n\mu)^2\right) \\
&= \frac{1}{n^2}\mathbb{E}\left(((X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu))^2\right)
\end{aligned}$$

where we have paired the n values of μ with the random variables $X_1, X_2, \ldots, X_n$. Observe that

$$((X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu))^2 = (X_1 - \mu)^2 + (X_2 - \mu)^2 + \cdots + (X_n - \mu)^2 + \sum_{j \neq k}(X_j - \mu)(X_k - \mu)$$

and since

$$\mathbb{E}((X_k - \mu)^2) = \text{Var}(X_k) = \sigma^2$$

for all k, the linearity of $\mathbb{E}$ guarantees that

$$\begin{aligned}\text{Var}(\overline{X_n}) &= \frac{1}{n^2}\mathbb{E}\left(((X-\mu)+(X_2-\mu)+\cdots+(X_n-\mu))^2\right)\\ &= \frac{1}{n^2}\left(n\sigma^2+\sum_{j\neq k}\mathbb{E}((X_j-\mu)(X_k-\mu))\right)\\ &= \frac{\sigma^2}{n}+\frac{1}{n^2}\sum_{j\neq k}\mathbb{E}((X_j-\mu)(X_k-\mu)).\end{aligned}$$

Using our hypothesis that the random variables $X_1, X_2, \ldots$ are independent and identically distributed, Proposition 6.20 ensures that

$$\mathbb{E}((X_j-\mu)(X_k-\mu)) = \mathbb{E}(X_j-\mu)\mathbb{E}(X_j-\mu) = (\mathbb{E}(X_k)-\mu)(\mathbb{E}(X_k)-\mu) = 0$$

whenever $j \neq k$. Thus,

$$\text{Var}(\overline{X_n}) = \frac{\sigma^2}{n}+\frac{1}{n^2}\sum_{j\neq k}0 = \frac{\sigma^2}{n}.$$

Finally, let $\epsilon > 0$. An appeal to the Bienaymé-Chebyshev inequality gives

$$\mathbb{P}\left(\left|\overline{X}_n-\mu\right|\geq\epsilon\right) \leq \frac{\text{Var}(\overline{X_n})}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{1}{n}\frac{\sigma^2}{\epsilon^2}$$

for every $n \in \mathbb{N}_+$. With this, the squeeze/sandwich theorem ensures that

$$0 \leq \lim_{n\to\infty}\mathbb{P}\left(\left|\overline{X_n}-\mu\right|\geq\epsilon\right) \leq \lim_{n\to\infty}\frac{1}{n}\frac{\sigma^2}{\epsilon^2} = 0$$

and so it follows that $\overline{X_n} \overset{\mathbb{P}}{\Rightarrow} \mu$. □

Example 7.6: The Frequentist Interpretation

Let Ω be a sample space equipped with probability $\mathbb{P}$ which represents a single experiment that we repeat over and over again independently ad infinitum thus producing and infinite number of independent trials. For an event $A \subseteq \Omega$, we saw in Example 7.2, that the number of occurrences in the first n trials is the random variable

$$S_n = X_1 + X_2 + \cdots + X_n$$

where the random variables $X_1, X_2, \ldots$ are independent and, for each k, X_k is a Bernoulli random variable with parameter $p = \mathbb{P}(A)$ giving the value 1 if A is observed on the kth trial and 0 if it isn't. Correspondingly, the random variables $X_1, X_2, \ldots$ are i.i.d Bernoulli random variables with common mean $\mu = \mathbb{E}(X_k) = \mathbb{P}(A)$ and finite variance. According to the Weak Law of Large Numbers, we have

$$\frac{\#(\text{Occurences of } A \text{ in first } n \text{ trials})}{n} = \overline{X_n}$$

converges to $\mu = \mathbb{P}(A)$ in probability. We note that "in probability" is in reference to the probability measure $\mathbb{P}_\infty$ on the sample space Ω^∞ of independent trials. In fact, more is true. As a consequence of the Strong Law of Large Numbers (Theorem 7.6 below),

$$\lim_{n\to\infty}\frac{\#(\text{Occurences of } A \text{ in first } n \text{ trials})}{n} = \mathbb{P}(A)$$

almost surely. That is, the event (in Ω^∞) has probability 1. This is exactly the frequentist interpretation of probability – which is how we were tempted to define probability measures in the first place. From a

statistical standpoint, if we perform an experiment over and over again in a way that we can ensure that each is done independently, the above tells us how to compute the probability of that event on the base sample space Ω.

The so-called Strong Law of Large Numbers, which we appealed to in the preceding example, is the following theorem:

Theorem 7.6 (The Strong Law of Large Numbers)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$. On Ω, let $X_1, X_2, X_3, \ldots$ be an i.i.d. sequence of random variables with common mean μ and variance $\sigma^2 < \infty$. Then*

$$\overline{X_n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{a.s.} \mu.$$

In other words, the statistical average of the random variables $X_1, X_2, \ldots$ converges to their common mean with probability 1.

Following the works of Bienaymé and Chebyshev, E. Borel is credited with the initial discovery of the result in the early twentieth century [12]; however, Borel's treatment only concerned binomial random variables and his proof contained a serious error. F. P. Cantelli generalized this result to general independent random variables with finite forth-order moments in 1917. Cantelli's argument was also found to be problematic and the question of to whom the theorem should be credited led to a serious and heated argument among mathematicians at the 1928 Congress of Mathematicians in Bologna. In addition to Borel and Cantelli, several other prominent mathematicians, including Khinchin, Slutsky, and Steinhaus, contributed to the result; notably, Slutsky was at the center of the argument in Bologna [12]. Though we give a satisfactory proof here assuming finite moments of order 4, most modern proofs of the theorem rely on an important result (appearing in both the works of Borel and Cantelli) which we now call the Borel-Cantelli Lemma. If you take a more advance course in probability, you will learn the Borel-Cantelli Lemma and its many consequences.

Proof. Needed. □

7.2 The Central Limit Theorem

Let $X_1, X_2, \ldots,$ be an i.i.d collection of random variables with common mean μ and variance $\sigma^2 < \infty$. Looking back at the weak law of large numbers, we see that, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(-n\epsilon < S_n - n\mu < n\epsilon\right) = \lim_{n\to\infty} \mathbb{P}\left(-\epsilon < \frac{S_n}{n} - \mu < \epsilon\right) = \lim_{n\to\infty} \mathbb{P}(\left|\overline{X_n} - \mu\right| < \epsilon) = 1.$$

In other words, this tells us the the probability that the sum S_n falls within $2n\epsilon$ of its mean $\mathbb{E}(S_n) = n\mu$ tends to 1 as $n \to \infty$. Though ϵ may be taken to be small, $n\epsilon$ grows without bound as $n \to \infty$ and so the above result tells us very little for predicting the value of sum S_n as $n \to \infty$. Thus, while the strong and weak laws of large numbers are extremely useful for saying something about the statistical average, they are much less useful for understanding the evolution of the sum S_n as $n \to \infty$. In the context of Alice's random walk, the law of large numbers says that the average of Alice's steps tends to $p - q$ as $n \to \infty$, however, it doesn't say much about Alice's location S_n for large n.

For the reason's discussed above, we'd like to have a better understanding of the sum S_n of an i.i.d. collection of random variables. A step in this direction is captured by the celebrated central limit theorem. give some history.

Theorem 7.7 (The Central Limit Theorem)**.** *Let Ω be a sample space equipped with probability measure $\mathbb{P}$ and let $X_1, X_2, \ldots$ be an i.i.d. collection of random variables with common mean μ and variance σ^2. Then*

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to a standard normal random variable Z. In particular, for any interval $I = (a, b)$,

$$\lim_{n\to\infty} \mathbb{P}\left(a < \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

More examples needed

We shall do our best to give some reasoning behind the proof of this theorem. Admittedly, our argument has a gap in that it relies on a characterization of convergence in distribution which is beyond the scope of these notes. Still, framing this characterization will lead us to the consideration of an important tool called the moment generating function; this tool has broad applications beyond our proof of the central limit theorem. For this reason, we spend the next subsection introducing it. Note Here

7.2.1 Moment Generating Functions

In this subsection, we discuss an important tool that aids us in the understanding of random variables. Though the following definition might seem "out of the blue", it gives us a helpful way to compute the moments of a random variable and will help us as we discuss convergence of random variables.

Definition 7.8. *Let X be a random variable on a sample space Ω equipped with probability measure $\mathbb{P}$. The moment generating function of X is the function real-valued function*

$$M_X(t) = \mathbb{E}(e^{tX})$$

which is defined for all real numbers t at which the expectation $\mathbb{E}(e^{tX})$ exists.

On a countable sample space, our definition of the expectation allows us to compute moment generating functions by summing over $\omega \in \Omega$. However, in practice, such computations are impractical and, instead, these computations are done (at least for discrete and continuous random variables) using theorems such as Theorem 5.41. Using this theorem, in particular, we immediately obtain the following.

Proposition 7.9. *Let X be a random variable on a sample space Ω equipped with probability measure $\mathbb{P}$.*

1. *If X is a discrete random variable with probability mass function p_X, then the moment generating function $M_X(t)$ is defined and given by*
$$M_X(t) = \sum_x e^{tx} p_X(x)$$
whenever $t \in \mathbb{R}$ is such that the above sum/series is finite.

2. *If X is a continuous random variables with probability density function f_X, then the moment generating function $M_X(t)$ is definted and given by*
$$M_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x)\, dx$$
whenever $t \in \mathbb{R}$ is such that the above integral converges (is finite).

Let's compute the moment generating functions of some familiar random variables.

Example 7.7: Bernoulli

Let $X \sim \text{Ber}(p)$ for $0 \leq p \leq 1$. Then

$$M_X(t) = \sum_x e^{tx} p_X(x) = e^{t\cdot 0} q + e^{t\cdot 1} p = q + pe^t$$

which is defined for all $t \in \mathbb{R}$.

Example 7.8: Binomial

Let $X \sim \text{Bin}(n, p)$ for $0 \leq p \leq 1$. We have

$$\begin{aligned} M_X(t) &= \sum_k e^{tk} p_X(t) \\ &= \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k q^{n-k} \\ &= (pe^t + q)^n \end{aligned}$$

which is defined for all $t \in \mathbb{R}$; here we have made use of the binomial theorem and the fact that $e^t k = (e^t)^k$.

Example 7.9: Geometric

Given a geometric random variable $X \sim \text{Geo}(p)$ with $0 < p \leq 1$, we have

$$\begin{aligned} M_X(t) &= \sum_{k=1}^{\infty} e^{tk} p q^{k-1} \\ &= \sum_{k=1}^{\infty} e^t e^{t(k-1)} p q^{k-1} \\ &= pe^t \sum_{k=1}^{\infty} (e^t q)^{k=1} \\ &= pe^t \frac{1}{1 - e^t q} \\ &= \frac{pe^t}{1 - e^t q} \end{aligned}$$

provided that $e^t q < 1$ or, equivalently, that $t < \log(1/q) = -\log(q)$.

Exercise 7.2: Poisson

Let $X \sim \text{Pois}(\lambda)$. Compute the moment generating function M_X of X.

Example 7.10: Exponential

Let $X \sim \text{Exp}(\lambda)$. We have

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx = \int_0^{\infty} e^{tx} \lambda e^{-\lambda x}\, dx = \int_0^{\infty} \lambda e^{(t-\lambda)x}\, dx.$$

In looking at the integrand, we expect that it might not converge when t is larger than λ. To sort this out,

let's view it, rightly, as an improper Riemann integral. For $t \neq \lambda$, we have

$$\begin{aligned}\lim_{s\to\infty}\int_0^s \lambda e^{(t-\lambda)x}\,dx &= \lim_{s\to\infty}\frac{\lambda}{t-\lambda}e^{(t-\lambda)x}\Big|_0^s \\ &= \lim_{s\to\infty}\frac{\lambda}{t-\lambda}\left(e^{(t-\lambda)s}-1\right) \\ &= \lim_{s\to\infty}\frac{\lambda}{\lambda-t}\left(1-e^{(t-\lambda)s}\right) \\ &= \frac{\lambda}{\lambda-t}\lim_{s\to\infty}\left(1-e^{(t-\lambda)s}\right)\end{aligned}$$

and this limit only exists when $t < \lambda$ (note, we have assumed that $t \neq \lambda$. In the case that $t = \lambda$,

$$\lim_{s\to\infty}\int_0^s \lambda e^{(t-\lambda)x}\,dx = \lim_{s\to\infty}\int_0^s \lambda\,dx = \infty.$$

Consequently, the moment generating function of X is defined only when $t < \lambda$ and, in this case, we have

$$M_X(t) = \int_0^\infty \lambda e^{(t-\lambda)x}\,dx = \lim_{s\to\infty}\int_0^s \lambda e^{(t-\lambda)x}\,dx = \frac{\lambda}{\lambda-t}$$

Example 7.11: Normal

In this example, we compute that moment generating function of a normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. To do this, as we've found with other properties of normal random variables, it is first helpful to work things out for a standard normal random variable Z. To this end, let $Z \sim \mathcal{N}(0,1)$ and observe that

$$M_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{tz}e^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{tz-z^2/2}\,dz.$$

Now, for real numbers t and z, we notice that

$$tz - z^2/2 = -\frac{1}{2}(z^2-2tz) = -\frac{1}{2}(z^2-2tz+t^2-t^2) = -\frac{(z-t)^2}{2}+\frac{t^2}{2}$$

and therefore

$$M_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(z-t)^2/2}e^{t^2/2}\,dz = \frac{e^{t^2/2}}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(z-t)^2/2}\,dz.$$

By making the change of variables $x = (z-t)$, we find that

$$\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-(z-t)^2/2}\,dz = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-x^2/2}\,dx = 1.$$

Thus, the moment generating function of a standard normal random variable Z is defined for all real numbers and is given by

$$M_Z(t) = e^{t^2/2}$$

for $t \in \mathbb{R}$. Now, in the case that $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $X = \sigma Z + \mu$ for $Z \sim \mathcal{N}(0,1)$ by virtue of Proposition 5.40. Therefore

$$M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left((e^{t(\sigma Z+\mu)}\right) = \mathbb{E}\left(e^{(\sigma t)Z}e^{\mu t}\right) = e^{\mu t}\mathbb{E}\left((e^{(\sigma t)Z}\right) = e^{\mu t}M_Z(\sigma t)$$

where we have used the fact that the expectation is linear and $e^{\mu t}$ is constant in the eyes of the expectation. Thus

$$M_X(t) = e^{\mu t}e^{(\sigma t)^2/2} = \exp((\sigma t)^2/2 + \mu t)$$

for $t \in \mathbb{R}$ where we are using the notation $\exp(z) = e^z$.

Exercise 7.3: Gamma

For $\alpha, \lambda > 0$, we recall that the Gamma density with parameters α, λ is given by

$$f(x) = \frac{\lambda}{\Gamma(\alpha)} e^{-\lambda x} (\lambda x)^{\alpha - 1}$$

for $x \geq 0$ and zero otherwise. For a so-called Gamma random variable X with $f_X = f$, show that

$$M_X(t) = \left(\frac{\lambda}{\lambda - t} \right)^{\alpha}$$

for $t < \lambda$ and does not exist otherwise. λ.

Now that we have a good handle on moment generating functions, it's high time to understand why they are important. One reason is given by the following result, a result which also explains the moniker "moment generating".

Proposition 7.10. *Let X be a random variable with moment generating function $M_X(t)$. If $M_X(t)$ is defined (i.e., finite) on an open interval containing 0, then X has finite absolute moments of all orders, the moment generating function is infinitely differentiable at $t = 0$ and, for each $n \in \mathbb{N}$,*

$$\mathbb{E}(X^n) = M_X^{(n)}(0) = \frac{d^n M_X}{dt^n}\Big|_{t=0}.$$

Proof. If $M_X(t)$ is finite on some open interval containing 0, then it must be finite on an interval of the form $[-t_0, t_0]$ for some $t_0 > 0$. Observe that

$$e^{t_0|x|} \leq e^{t_0 x} + e^{-t_0 x}$$

for all real numbers x. By the monotonicity of the expectation, it follows that

$$\mathbb{E}(e^{t_0|X|}) \leq \mathbb{E}(e^{t_0 X} + e^{-t_0 X}) = \mathbb{E}(e^{t_0 X}) + \mathbb{E}(e^{-t_0 X}) = M_X(t_0) + M_X(-t_0) < \infty.$$

Consequently,

$$\sum_{n=0}^{\infty} \frac{t_0^n \mathbb{E}(|X|^n)}{n!} = \sum_{n=0}^{\infty} \mathbb{E}\left(\frac{(t_0 |X|)^n}{n!} \right) = \mathbb{E}\left(\sum_{n=0}^{\infty} \frac{(t_0 |X|)^n}{n!} \right) = \mathbb{E}(e^{t_0|X|}) < \infty;$$

here, we used the Maclaurin expansion $e^z = \sum_{n=0}^{\infty} z^n/n!$ for the exponential function and the reason that we were able to exchange the expectation and the series follows from the so-called monotone convergence theorem of measure theoryNeed citation. In particular, this shows that $\mathbb{E}(|X|^n) < \infty$ for each n and

$$\lim_{N \to \infty} \sum_{n=N+1}^{\infty} \frac{t_0^n \mathbb{E}(|X|^n)}{n!} = 0. \tag{7.2}$$

I claim that

$$M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}(X^n) \tag{7.3}$$

for all $-t_0 \leq t \leq t_0$. Using properties of power series (centered at zero), the validity of this identity would guarantee, at once, that $M_X(t)$ is differentiable on the interval $(-t_0, t_0)$ and that $M_X^{(n)}(0) = \mathbb{E}(X^n)$ for every natural number n (you should check this!).

To establish the identity (7.3), observe that, for each $N \in \mathbb{N}$ and $-t_0 \leq t \leq t_0$,

$$\begin{aligned}\left|M_X(t) - \sum_{n=0}^{N} \frac{t^n}{n!}\mathbb{E}(X^n)\right| &= \left|\mathbb{E}(e^{tX}) - \mathbb{E}\left(\sum_{n=0}^{N}\frac{(tX)^n}{n!}\right)\right| \\ &= \left|\mathbb{E}\left(\sum_{n=N+1}^{\infty}\frac{(tX)^n}{n!}\right)\right| \\ &\leq \mathbb{E}\left(\left|\sum_{n=N+1}^{\infty}\frac{tX^n}{n!}\right|\right) \\ &\leq \mathbb{E}\left(\sum_{n=N+1}^{\infty}\frac{|tX|^n}{n!}\right) \\ &\leq \mathbb{E}\left(\sum_{n=N+1}^{\infty}\frac{t_0\,|X|^n}{n!}\right) \\ &= \sum_{n=N+1}^{\infty}\frac{t_0^n}{n!}\mathbb{E}(|X|^n)\end{aligned}$$

where we have used the monotonicity of expectation, the triangle inequality for the infinite summation, and the monotone convergence theorem once again. Thus, for every $-t_0 \leq t \leq t_0$, it follows from (7.2) that

$$\left|M_X(t) - \sum_{n=0}^{\infty}\frac{t^n}{n!}\mathbb{E}(X^n)\right| = \lim_{N\to\infty}\left|M_X(t) - \sum_{n=0}^{N}\frac{t^n}{n!}\mathbb{E}(X^n)\right| \leq \lim_{N\to\infty}\sum_{n=N+1}^{\infty}\frac{t_0^n}{n!}\mathbb{E}(|X|^n) = 0$$

or, equivalently, (7.3) holds for each $-t_0 \leq t \leq t_0$. □

Example 7.12: Moments of the Binomial Random Variable

Given that the probability mass function of a binomial random variable $X \sim \text{Bin}(n,p)$ is simply a finite sum, all of its (absolute) moments are finite and so M_X is infinitely differentiable at $t = 0$ and we have

$$\mathbb{E}(X^k) = M_X^{(k)}(0)$$

for all $k \in \mathbb{N}$. We recall from Example 7.2 that $M_X(t) = (q + pe^t)^n$ for $t \in \mathbb{R}$ and so, in particular,

$$\mathbb{E}(X) = \frac{d}{dt}(q+pe^t)^n\Big|_{t=0} = npe^t(q+pe^t)^{n-1}\Big|_{t=0} = npe^0(q+p)^{n-1} = np.$$

This is, of course, what we already knew. However, notice how much easier it was to compute.

Exercise 7.4: Computing Moments

Use the proposition above to compute the means of the Geometric, Poisson, Exponential, and Normal Random variables.

Another important property of moment generating functions is captured by the following result.

Proposition 7.11. *Let X and Y be independent random variables with moment generating functions $M_X(t)$ and M_Y respectively. Then the moment generating function of $X + Y$ is given by*

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

and is defined precisely for those values of t at which both M_X and M_Y are defined.

Proof. Writing $Z = X + Y$, we have

$$M_Z(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX}e^{tY}).$$

Since X and Y are independent, Proposition 6.20 guarantees that

$$\mathbb{E}(e^{tX}e^{tY}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY})$$

and therefore

$$M_Z(t) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}) = M_X(t)M_Y(t).$$

□

As an immediate corollary, we have the following:

Corollary 7.12. *Let $X_1, X_2, X_3, \ldots$ be an independent and identically distributed sequence of random variables with (necessarily) common moment generating function $M = M_{X_k}$ for $k = 1, 2, \ldots$. For $n \in \mathbb{N}_+$, define*

$$S_n = X_1 + X_2 + \cdots + X_n.$$

Then,

$$M_{S_n}(t) = M(t)^n.$$

Example 7.13:

To illustrate the corollary above, let $X_1, X_2, \ldots, X_n$ be Bernoulli random variables with common parameter p. We found that

$$M(t) = M_{X_k}(t) = q + pe^t$$

for $k = 1, 2, \ldots, n$ and $t \in \mathbb{R}$. In view of the corollary above, we have

$$M_{S_n}(t) = M(t)^n = (q + pe^t)^n$$

for the random variable $S_n = X_1 + X_2 + \cdots + X_n$. This comes as no surprise for, in looking back at Example 6.19 (and Proposition 6.25 in particular), we found that the sum of n i.i.d Bernoulli random variables is a binomial random variable with parameters n and p.

We finish this subsection by presenting a characterization of convergence in distribution in terms of moment generating functions. The proof of this characterization is beyond the scope of these notes. It relies on an analogous one given in terms of characteristic functions which, itself, relies on the famous continuity theorem of P. Levý. [4].

Proposition 7.13. *Let $X_1, X_2, \ldots,$ be a sequence of random variables with moment generating functions $M_1 = M_{X_1}, M_2 = M_{X_2}, \ldots$. Also, let X be another random variable with moment generating function $M = M_X$. If*

$$\lim_{n\to\infty} M_n(t) = M(t)$$

for all t sufficiently close to 0, then the sequence X_n converges in distribution to X.

Exercise 7.5:

Earlier this semester, we discussed how binomial random variables could be modeled by Poisson random variables where p was small, n was large and $\lambda = np$ was fixed. In this exercise, you will use the preceding theory to make precise this earlier discussion. To this end, let $X_n \sim \text{Bi}(n, p) = \text{Bi}(n, \lambda/n)$ where $\lambda > 0$ is fixed and denote by $M_n(t)$ the moment generating function of X_n.

1. Compute $M_n(t)$.

2. Show that
$$\lim_{n\to\infty} M_n(t) = \exp(\lambda(e^t - 1))$$
and use it to conclude (by virtue of the proposition) that X_n converges to a Poisson random variable with parameter λ in distribution.

Exercise 7.6: Alice's Random Walk

For $k = 1, 2, \ldots$, let X_k be a random variable taking the values -1 and 1 each with probability $1/2$. Assume that the random variables $X_1, X_2, \ldots$ are (mutually) independent and, for each n, define
$$S_n = X_1 + X_2 + \cdots X_n.$$

1. Compute $M_{S_n}(t)$.
2. Show that
$$\lim_{n\to\infty} M_{S_n/\sqrt{n}}(t) = e^{t^2/2}.$$
3. Using Proposition 7.13, conclude that
$$\frac{S_n}{\sqrt{n}} = \frac{X_1 + X_2 + \cdots X_n}{\sqrt{n}}$$
converges in distribution to the standard normal distribution. Use this to show that, for each $\epsilon > 0$,
$$\lim_{n\to\infty} P\left(-\epsilon \leq \frac{S_{2n}}{\sqrt{2n}} \leq \epsilon\right) = \int_{-\epsilon}^{\epsilon} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \approx \frac{2\epsilon}{\sqrt{2\pi}}$$
where the final approximation is valid only when ϵ is small.
4. Finally, by setting $\epsilon = (2\sqrt{2n})^{-1}$, confirm that this approximation (formally, not rigorously) says that
$$P(S_{2n} = 0) = P(-\frac{1}{2} \leq S_{2n} \leq \frac{1}{2}) \approx \frac{1}{\sqrt{4\pi n}}.$$
If you look back through your notes, we got precisely this answer using Stirling's formula (and many more assumptions)!

7.2.2 The proof of the central limit theorem

Proof of Theorem 7.7. Let $X_1, X_2, \ldots$ be i.i.d random variables with mean μ and finite variance σ^2. We shall assume additionally that the common moment generating functions of the random variables $X_1, X_2, \ldots$ is finite on an interval containing 0 (so that we may use Proposition 7.10). For each n, define
$$X_n' = \frac{X_n - \mu}{\sigma}$$
and observe that $\mathbb{E}(X_n') = (\mathbb{E}(X_n) - \mu)/\sigma = 0$ and
$$\text{Var}(X_n') = E((X_n')^2) = \frac{1}{\sigma^2}\mathbb{E}((X_n - \mu)^2) = \frac{\text{Var}(X_n)}{\sigma^2} = 1.$$
Thus, our new necessarily i.i.d. sequence of random variables (do you see why they are i.i.d) $X_1', X_2', \ldots$ all have zero mean and unit variance. Let us denote M the common moment generating function of the random variables

$X_1', X_2', \ldots$. We note that, given our assumption that the common moment generating function of our (original) random variables $X_1, X_2, \ldots$ is finite on an interval containing 0, it follows straightforwardly (why?) that $M(t)$ is also finite on a neighborhood of 0 and so, in particular, $M(t)$ is infinitely differentiable on that interval (with every derivative continuous) and $M(0) = 1$, $M'(0) = \mathbb{E}(X_k') = 0$, and $M''(0) = \mathbb{E}((X_k')^2) = \text{Var}(X_k') = 1$, for every k.

Let

$$S_n' = X_1' + X_2' + \cdots + X_n'$$

and observe that

$$S_n' = \frac{(X_1 - \mu)}{\sigma} + \frac{(X_2 - \mu)}{\sigma} + \cdots + \frac{(X_n - \mu)}{\sigma} = \frac{S_n - n\mu}{\sigma}$$

for each n. Correspondingly, the cumulative distribution of $S_n'/\sqrt{n}$ is coincides with that of $(S_n - n\mu)/\sqrt{n}\sigma$ and, in view of Proposition 7.13, it suffices to prove that

$$\lim_{n\to\infty} M_{S_n'/\sqrt{n}}(t) = M_Z(t) = e^{t^2/2} \tag{7.4}$$

since $M_Z(t) = e^{t^2/2}$ is the moment generating function of a standard normal random variable Z. To establish (7.4), observe that

$$M_{S_n'/\sqrt{n}}(t) = \mathbb{E}(e^{tS_n'/\sqrt{n}}) = \mathbb{E}(e^{(t/\sqrt{n})S_n'}) = M_{S_n'}(t/\sqrt{n}) = (M(t/\sqrt{n}))^n.$$

where we have made use of Corollary 7.12 to see that the moment generating function of S_n' coincides with the nth power of M. Using two applications of L'Hôpital's rule, we have

$$\begin{aligned}
\lim_{n\to\infty} n \log M(t/\sqrt{n}) &= \lim_{x\to 0} \frac{\log(M(tx))}{x^2} \\
&= \lim_{x\to 0} \frac{tM'(tx)}{2xM(tx)} \\
&= \lim_{x\to\infty} \frac{t^2 M''(tx)}{2M(tx) + 2xM'(tx)} \\
&= \lim_{x\to\infty} \frac{t^2 M''(0)}{2M(0)} \\
&= t^2/2
\end{aligned}$$

where we have used the continuity of M, M' and M'' and the fact that $M(0) = M''(0) = 1$ and $M'(0) = 0$. Therefore

$$\lim_{n\to\infty} M_{S_n'/\sqrt{n}}(t) = \lim_{n\to\infty} \exp(\log(M(t/\sqrt{n})^n)) = \lim_{n\to\infty} \exp(n \log M(t/\sqrt{n})) = \exp(t^2/2)$$

where we have used the continuity of the exponential function (and the fact that, for any $a > 0$, $a^n = e^{n\log(a)}$). Thus we have established 7.4 and our theorem is proved. □

Appendix A

A.1 Some Calculus Facts

Proposition A.1. *Let $A = \{a_n\}$ be a sequence of numbers with $\lim_{n\to\infty} a_n = L$. Then*

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^L.$$

Proof. I will first prove a special case of the proposition where the sequence is identically constant, i.e., $a_n = L$ for all n. In this case, a standard argument using L'Hôpital's rule is applicable (and so the proof is easy to understand at the level of single-variable calculus). A completely different proof of the general situation will be given after; this proof makes use of a big theorem called the Dominated Convergence Theorem and is really beyond the scope of these notes. Still, I am including this proof for the interested reader.

For the sequence $a_n = L$ for all n, we have

$$\left(1 + \frac{a_n}{n}\right)^n = \exp\left(\log\left[\left(1 + \frac{L}{n}\right)^n\right]\right) = \exp\left(n \log\left(1 + \frac{L}{n}\right)\right)$$

whenever n is large enough so that $1 + L/n > 0$. Now,

$$\lim_{n\to\infty} n \log\left(1 + \frac{L}{n}\right) = \lim_{n\to\infty} \frac{\log\left(1 + \frac{L}{n}\right)}{1/n} = \lim_{t\to 0} \frac{\log(1 + Lt)}{t}.$$

Since this last limit contains a so-called indeterminate form (and both the numerator and denominator are continuously differentiable), we apply L'Hôpital's rule to see that

$$\lim_{n\to\infty} n \log\left(1 + \frac{L}{n}\right) = \lim_{t\to 0} \frac{\frac{d}{dt}\log(1 + Lt)}{\frac{d}{dt}t} = \lim_{t\to 0} \frac{\frac{L}{1+Lt}}{1} = L.$$

Thus, by the continuity of the exponential function, we find that

$$\lim_{n\to\infty} \left(1 + \frac{L}{n}\right)^n = \lim_{n\to\infty} \exp\left(n \log\left(1 + \frac{L}{n}\right)\right) = \exp\left(\lim_{n\to\infty} n \log\left(1 + \frac{L}{n}\right)\right) = \exp(L) = e^L.$$

Thus, for the constant sequence $a_n = L$, we have proven the proposition.

Let's now prove the result in general. Since $A = \{a_n\}$ is a convergent sequence, it is bounded and so we are able to find some M for which $|a_n| \leq M$ for all n. Using the binomial theorem, we have

$$\begin{aligned}
\left(1 + \frac{a_n}{n}\right)^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{a_n^k}{n^k} \\
&= \sum_{k=0}^{\infty} \binom{n}{k} \frac{a_n^k}{n^k} \mathbb{1}_{[0,n]}(k) \\
&= \sum_{k=0}^{\infty} g_n(k)
\end{aligned}$$

where

$$g_n(k) = \binom{n}{k} \frac{a_n^k}{n^k} \mathbb{1}_{[0,n]}(k)$$

for $k, n \in \mathbb{N}$. Observe that

$$\begin{aligned} g_n(k) &= \frac{n!}{(n-k)!n^k} \frac{a_n^k}{k!} \mathbb{1}_{[0,n]}(k) \\ &= \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} \frac{a_n^k}{k!} \mathbb{1}_{[0,n]}(k) \\ &= \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right) \frac{a_n^k}{k!} \mathbb{1}_{[0,n]}(k) \end{aligned}$$

and from this we seen that, for each $k \in \mathbb{N}$,

$$|g_n(k)| \leq 1 \cdot 1 \cdots 1 \frac{|a_n|^k}{k!} \cdot 1 \leq \frac{M^k}{k!} =: g(k)$$

and

$$\lim_{n\to\infty} g_n(k) = \frac{L^k}{k!}$$

since $\lim_{n\to\infty} a_n^k = L^k$ and $\lim_{n\to\infty} \mathbb{1}_{[0,n]}(k) = 1$. Since $g(k)$ is summable, i.e.,

$$\sum_{k=0}^{\infty} g(k) = \sum_{k=0}^{\infty} \frac{M}{k!} = e^M < \infty,$$

the dominated convergence theorem [1, Theorem 5.6] (applied to the Lebesgue integral using counting measure), we have

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = \lim_{n\to\infty} \sum_{k=0}^{\infty} g_n(k) = \sum_{k=0}^{\infty} \lim_{n\to\infty} g_n(k) = \sum_{k=0}^{\infty} \frac{L^k}{k!} = e^L,$$

as was asserted. □

Bibliography

[1] R. Bartle. The Elements of Integration and Lebesgue Measure. (1966)

[2] Bennett, T. "How many particles are in the observable universe." Popular Mechanics online (2017).

[3] P. Billingsley. Probability and Measure, 3rd Ed. John Wiley and Sons, Inc. (1995).

[4] J. H. Curtiss. "A Note on the Theory of Moment Generating Functions." Annals of Mathematical Statistics, 13(4):430-433, 1942.

[5] R. Durrett. "Triple birthday matches in the Senate: Lies, damned lies, and chatGPT" arXiv preprint arXiv:2302.09643 (2023). https://doi.org/10.48550/arXiv.2302.09643

[6] M. Martinez-Bakker M, K. M. Bakker, A. A. King, P. Rohani."Human birth seasonality: latitudinal gradient and interplay with childhood disease dynamics." Proc. R. Soc B 281: 20332438. (2014) http://dx.doi.org/10.1098/rspb.2013.2438

[7] Moore, Calvin C, "Ergodic theorem, ergodic theory, and statistical mechanics.", PNAS, 112(7):1907-1911, 2015.

[8] Stein, Elias M. and Shakarchi, Rami. Real analysis: measure theory, integration, and Hilbert spaces. Princeton University Press. Princeton University Press. (2005)

[9] G. R. Grimmett and D. R. Stirzaker. Probability and Random Processes, 2nd Ed. Oxford University Press (1992)

[10] M. M. Rao and R. J. Swift. Probability Theory with Applications. 2nd. Ed., Springer (2006)

[11] Sheldon Ross. A First Course in Probability. Which edition?

[12] E. Seneta, "On the History of the Strong Law of Large Numbers and Boole's Inequality." Historia Mathematica, 19:24-39, 1992.

[13] E. Seneta, "A Tricentenary history of the Law of Large Numbers." Bernoulli, 19(4):1088-1121, 2013.

[14] Kai Lai Chung and Farid AitSahlia. Elementary Probability Theory, with Stochastic Processes and an Introduction to Mathematical Finance, 4th Ed. Springer-Verlag, New York (2003)

[15] Rudin, Walter. Principles of Mathematical Analysis. McGraw-Hill Inc. New York. (1976)