# Introduction to Probability & Statistics

Charles C. Johnson

December 10, 2018

# Contents

# *Introduction to the Course*

*Difficulties strengthen the mind, as labor does the body.*

SENECA THE YOUNGER

## Overview

Welcome to Math M-365, the first course in probability and statistics at Indiana University. The goal of this course is to teach you the fundamentals of probability theory and random variables, and how these ideas are applied to inferential statistics. The course is divided up into three main portions, each of which should roughly take about one third of the semester and more-or-less corresponds to each exam during the regular semester (i.e., before the final exam).

Before getting started on the three main portions of the course, though, we will have a very quick introduction/review to set theory, which is the basic language of most modern mathematics. You have likely seen parts of this material before in other courses, but to make sure everyone is on the same page we will start from the very basics. Because of time constraints we will likely not spend more than the first week of class discussing this material in lecture, but the details of everything discussed in class will be fleshed out in these lecture notes.

After the set theory introduction we will discuss probability theory. We will begin by defining the ideas of experiments, sample spaces, events, and probability very precisely. After the basic definitions we will discuss some combinatorial (aka counting) techniques which will be helpful for solving problems where we need to determine all the possible outcomes of an experiment. We will then move on to discussing conditional probability which tells us how partial information about the outcome of an experiment can help us compute probabilities, and finally we will discuss Bayes' theorem and related topics.

In the second third of material we will discuss random variables, which are functions defined on the set of outcomes of a random experiment. We will see that there are two basic types of random variables, called discrete and continuous, and will also spend some time discussing some important families of each type of random variable. Initially we will only concern

ourselves with one random variable at a time, but later we will be interested in dealing with several random variables simultaneously. This will require a brief excursion into the technique of iterated integrals from calculus, but we will define everything we need in class for the benefit of anyone that hasn't seen (or has seen and forgotten) that material.

Finally, we will turn our attention to statistics. Here we will apply the theory, tools, and techniques we've developed while discussing probability and random variables to study how we can infer information about an entire population based on a sample of individuals from that population. In particular, we will discuss point estimators and confidence intervals, which give us tools for estimating parameters of the population from sample data. We will also discuss hypothesis testing which can tell us if there is enough evidence to accept or reject a claim about the population.

Many of the ideas we will discuss in this class have direct application to a variety of real-world problems, and when time permits we will discuss applications of the material. Most of the interesting applications will have to wait until we've developed some theory, however.

## Prerequisites

This course is meant to be a first introduction to probability theory and statistics, and assumes no previous knowledge of either of these topics. Having said that, there is a bit of "mathematical maturity" that is assumed. Officially this means the prerequisites for this course are calculus, and it assumed that you are familiar with basic ideas and techniques learned in the first two semesters of the calculus sequence (e.g., the various derivative rules, integration formulas, and the main theorems from calculus). Any other mathematical background not covered in the typical calculus sequence will be developed in class as needed.

## How to do well in this course

For most students, this course will be more demanding and require more work than most of their previous math courses. Though the course will start from basic principles and require relatively simple computations initially, by the end of the semester we will be using some fairly sophisticated ideas and techniques on a regular basis. The material in this course is by its very nature cumulative: everything we do in class will build off of previous material. For this reason it is important that you stay up-to-date with the

material discussed in class. Because of the pace of the class, falling behind will make it extremely difficult to catch up. You need to understand this so that you can be prepared to devote time to studying for the course on a regular and consistent basis.

My recommendation is that you come to class every class period whenever possible; take notes during class; shortly after class review your notes; read the lecture notes online; read the textbook for the course; and do as many practice problems as you can. Practice problems will be scattered throughout the lecture notes, and solutions to the problems will appear at the end of the notes. Many of the practice problems are actual problems from previous homeworks, quizzes, and exams and should help you to prepare for your own assignments.

Some of the topics we discuss in this class will be confusing at first, and it's okay if you don't understand everything at first. The important thing is to continually work hard, think about the material, and ask questions if you don't understand something. This can be frustrating and time-consuming, but it is the only way to learn some of the difficult material you will encounter in this course.

Be sure to start homework assignments early, as some problems can be tricky and require a little bit of time to figure out. This won't be an issue *if* you start the assignment early and try to do a few problems each day. However, if you wait until the day before an assignment is due to start it, it's unlikely you will be able to finish before the due date.

Even though the material in this course can be challenging, you can master it if you're willing to work hard and not allow any initial frustrations to prevent you from continuing to study.

# Part I

# Preliminaries

# 1 Naive Set Theory

*A set is a Many which allows itself to be thought of as a One.*

GEORG CANTOR

Before we can discuss probability theory, we need to set up some basic ideas from *set theory*, although we will do this in a somewhat hand-wavy way. Set theory provides a foundation for most of mathematics, even though this point of view often isn't emphasized in more basic courses. For our purposes in this course, we will primarily treat set theory as a convenient language for organizing ideas.

Some of the proofs of the facts we state in this chapter, as well as in chapters two and three, are a little technical and will probably seem confusing if this is your first time learning this material. Don't let this worry you too much: the proofs are included mostly for completeness and you can safely skip reading them if you want. You should, however, know all of the definitions and statements of theorems since we will use them later in the semester.

## 1.1 Sets

### Definitions and examples

A ***set*** is an unordered collection of objects. These objects could be numbers, points in space, functions, words, symbols, other sets, or (almost) anything else. Most of mathematics is described in terms of sets, even though this isn't always made explicit.

We sometimes describe a set by explicitly writing out everything in the set, separated by commas, and surrounded by curly braces. For example, the set containing the first few positive, even numbers is

$$\{2, 4, 6, 8, 10\}.$$

The only thing that matters when we talk about a set is what is in the set. The order in which an object occurs in a set does not matter, so the following sets are all the same:

$$\{2, 4, 6, 8, 10\} = \{10, 8, 6, 4, 2\} = \{8, 2, 4, 10, 6\}.$$

The number of times we write an object in the set also does not matter (as long as it occurs at least once):

$$\{2, 4, 6, 8, 10\} = \{2, 2, 2, 4, 4, 6, 8, 10, 10, 10, 10, 10\}.$$

We use the symbol $\in$ to denote that something is an element of a set, and $\notin$ to denote that something is not an element of a set:

$$2 \in \{2, 4, 6, 8, 10\}$$
$$3 \notin \{2, 4, 6, 8, 10\}.$$

Many times a set will be too big for us to write out all of the elements, and in that situation we need some other notation to describe the set. One common notation is to list a few elements in a set and then write "..." to mean "continue the pattern." For example,

$$\{2, 4, 6, 8, 10, 12, 14, ...\}$$

denotes the set of all positive even numbers; while

$$\{5, 10, 15, 20, 25, 30, ...\}$$

denotes the set of all positive multiples of 5.

**Exercise 1.1.**
Write down a set which contains all positive integers that satisfy the following conditions: each number is a multiple of 4, a multiple of 6, and is less than 50.

Of course, it can get tedious to write sets down in this way every time we want to refer to a set. To save ourselves some writing if we are going to refer to a set multiple times, we will often assign the set a name. For example, if we write

$$E = \{2, 4, 6, 8, 10, 12, 14, ...\}$$

then we are saying we want to use the symbol $E$ to refer to the set of all positive even numbers. We are then justified in writing things like $28 \in E$, $17 \notin E$, and $-2 \notin E$.

It will sometimes be convenient to say that several things are or are not in a given set. In this case we list all of those things separated by commas and followed by $\in$ or $\notin$:

$$8, 32, 96, 384 \in E$$
$$3, 347, -10 \notin E$$

Many times the sets we will be interested in will be "special," and we will only be interested in those sets for a little while – e.g., while we're solving a particular problem. So, we might use $E$ to denote one set now and then later use the same symbol again to denote a different set. For instance, in solving one problem we may let $E$ denote the set $\{1, 2, 3\}$, and let we'll use $E$ to denote the set $\{-3, 7, 8\}$. It will usually be clear from context which set a given symbol refers to.

There are some sets that are used over and over, again and again, and those sets have special names and symbols that are reserved only for those particular sets. One such set is the set of ***natural numbers***, which is the set of all positive whole numbers and is denoted by a capital $N$, but written in what is often called "blackboard bold" and looks like $\mathbb{N}$:

$$\mathbb{N} = \{1, 2, 3, 4, 5, 6, ...\}.$$

**Remark.**
In older textbooks this $\mathbb{N}$ was originally written as a bold **N**. It's difficult to write bold letters on paper or a blackboard, however, and so people started writing an extra line in the letter to denote the letter was bold. This way of writing bold letters eventually became popular enough that it made its way into typed works as a special typeface.

The set of all whole numbers (positive, negative, and zero) is called the set of ***integers*** and is denoted by a blackboard bold $\mathbb{Z}$:

$$\mathbb{Z} = \{..., -3, -2, -1, 0, 1, 2, 3, ...\}.$$

**Remark.**
Using the letter $\mathbb{Z}$ might seem like a weird choice for integers, but it's only weird if you're an English speaker. Many influential mathematicians of the past, including Georg Cantor who is considered the father of set theory, were German and so they of course used the German equivalent of these words and used the first letter of those German words. The German word for numbers is *die Zahlen* (*die* is the feminine definite article in German, like *la* in French or Spanish), hence the $\mathbb{Z}$.

Conveniently, German and English have some commonalities and so some German words are very similar to their English counterparts, so most of these blackboard bold letters are actually what you would guess using the English words. For example, *the natural numbers* in German is *die natürliche Zahlen*, so $\mathbb{N}$ makes sense in both German and English.

The number of distinct elements in a set $A$ is called the ***cardinality*** of the set and is denoted by either $\#A$ or $|A|$. For example $\#\{7, 8, 0, 4, 3\} = 5$ while $\#\{3, 6, 9, 12, ..., 84, 87, 90\} = 30$. The cardinality can be infinite as well; both $\mathbb{N}$ and $\mathbb{Z}$ have infinite cardinality.

## 1.2 Set-builder notation

Unfortunately, there are times when the ... notation mentioned above can be ambiguous. For example,

$$\{2, 4, ...\}$$

could mean the set of all even numbers, or it could be all the powers of 2: both of the following sets match the pattern

$$\{2, 4, 6, 8, 10, ...\}$$
$$\{2, 4, 8, 16, 32, ...\}.$$

To get around this ambiguity we sometimes use ***set builder notation***. In this notation we write two curly braces, like normal, but separated into two parts by a vertical bar. On the left-hand side of the bar we write a variable (or sometimes a collection of variables) that give us some pattern that all of the elements in the set follow, and on the right-hand side we given a condition (usually in the form of an equation or inequality, but sometimes

written in words) that the variable must satisfy in order to be an element of the set. The collection of all positive even integers, for example may be written in set builder notation as

$$E = \{x \mid x = 2n \text{ for some } n \in \mathbb{N}\}.$$

That is, we start off by considering the natural numbers, but to be an element of $E$, a given natural number $x$ has to be two times some other natural number. (A number is even if and only if it is divisible by two.)

We could define the set of all positive odd numbers as

$$\mathcal{O} = \{x \mid x = 2n - 1 \text{ for some } n \in \mathbb{N}\}.$$

**Exercise 1.2.**

(a) Write the set of all positive, whole number multiples of 5 in set builder notation.

(b) Write the set of all whole number multiples (including negatives) of 5 in set builder notation.

Another common set of numbers is the set of ***rational numbers***, which are ratios of integers where the denominator is not zero. These are quotients[1], so the set of all rational numbers is denoted $\mathbb{Q}$. In set builder notation we can express $\mathbb{Q}$ as

$$\mathbb{Q} = \left\{ \frac{p}{q} \,\middle|\, p, q \in \mathbb{Z} \text{ and } q \neq 0 \right\}.$$

In the examples thus far we have only considered sets of numbers, but there is nothing special about numbers: the elements of a set can be any type of object. They could be names of people,

$$\{\text{William, Charles, Percy, Fred, George, Ron, Ginny}\},$$

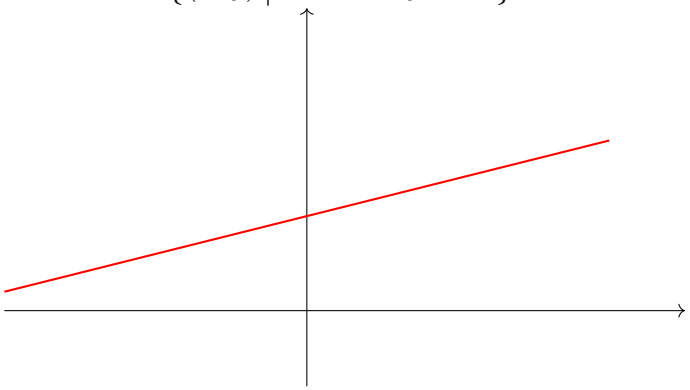
or abstract symbols,

$$\{\heartsuit, \clubsuit, \diamondsuit, \spadesuit\},$$

or points in space,

[1] Conveniently, the German word for *the quotient* is *der Quotient*, and so the $\mathbb{Q}$ makes sense for English speakers too!

$$\big\{(x,y)\,\big|\,-2x+8y=10\big\}.$$



You can even have sets that contain other sets:

$$\{\{1,2\},\{1,3\},\{2,3\}\}.$$

Sets are ubiquitous in mathematics: the vast majority of things you work with are, or are defined in terms of, sets. This point may not have been made clear to you before in earlier mathematics courses because it may not have been needed, but for our purposes in this class we will need to deal with sets on a regular basis, so it's important that we have a good understanding of them.

## 1.3 Subsets and supersets

We say that a set $A$ is a ***subset*** of a set $B$ if every element of $A$ is also an element of $B$. When this happens we write $A \subset B$.

**Example 1.1.**
Every natural number is an integer, so the set of natural numbers is a subset of the set of integers: $\mathbb{N} \subset \mathbb{Z}$. Every integer is also a rational number (e.g., $3 = {}^{3}/_{1}$), so $\mathbb{Z} \subset \mathbb{Q}$.

**Example 1.2.**
Suppose that $A$ is the set of all the multiples of 3, and $B$ is the set of

all multiples of 12:

$$A = \{x | x = 3n \text{ for some } n \in \mathbb{N}\},$$
$$B = \{x | x = 12n \text{ for some } n \in \mathbb{N}\}.$$

Since every multiple of 12 is also a multiple of 3 (because 3 divides 12), $B$ is a subset of $A$: $B \subset A$.

When $A$ is a subset of $B$ we say that $B$ is a ***superset*** of $A$. That is, when we write $A \subset B$ the set on the left is a subset of the set on the right; and the set on the right is a superset of the set on the left. The superset is the "larger" set, and the subset is the "smaller" set. Sometimes it will be convenient for the symbol $\subset$ to be written in the other direction: for example $B \supset A$. Here $B$ is still the larger superset, and $A$ is the smaller subset. (Compare this to writing $3 < 4$ and $4 > 3$.)

In our mind's eye we often picture the relationship between a set and a any subsets or supersets as shown in Figure 1.1.

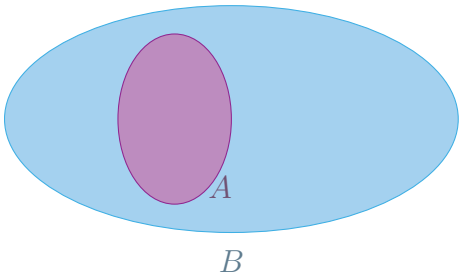



Figure 1.1: If $A \subset B$, then we imagine $A$ as sitting inside of $B$.

We will use pictures like this, which are called ***Venn diagrams***, many times when describing sets, even if the sets we're talking about don't really look like the two-dimensional shapes we'll draw: though the pictures aren't technically accurate (e.g., $A$ and $B$ may not be actually be the set of points making up two ovals in the plane), it's often very helpful to use these kinds of abstract pictures because they provide us with some intuition about how different sets are related to one another.

Notice again that we say $A$ is a subset of $B$ if every element of $A$ is also an element of $B$. This means, in particular, that for every set $A$, $A$ is a subset of itself: every element of $A$ is also an element of $A$. We are thus justified in writing $A \subset A$. If we want to explicitly exclude this possibility,

we use the symbol $\subsetneq$: writing $A \subsetneq B$ means that $A$ is a subset of $B$ and $A$ is not all of $B$.

**Exercise 1.3.**
Suppose that $A$ and $B$ are two sets and $A \subsetneq B$. Show that this means there must exist at least one element of $B$ which is not an element of $A$.

When $A \subsetneq B$ we call $A$ a ***proper subset*** of $B$. For example, the natural numbers are a proper subset of the integers, and the integers are a proper subset of the rational numbers.

**Remark.**
There is a little bit of ambiguity that can occur with the symbol $\subset$: some authors use $\subset$ to mean $\subsetneq$, and use $\subseteq$ to mean $\subset$. That is, some people will use $A \subseteq B$ to mean that $A$ is a subset of $B$, possibly all of $B$, and $A \subset B$ to mean that $A$ is a subset of $B$ but not all of $B$. This is reminiscent to using $\leq$ and $<$ in comparing numbers, but it's not completely standard.

To avoid any potential ambiguity we will typically use $A \subsetneq B$ to mean that $A$ is a proper subset of $B$, and $A \subseteq B$ to mean that $A$ is a subset of $B$ but could potentially be all of $B$.

## 1.4 Equality

We say that two sets $A$ and $B$ are ***equal*** if they have precisely the same elements: that is, if $x \in A$ then $x \in B$ and if $y \in B$, then $y \in A$ as well. This is exactly the same thing as saying $A \subseteq B$ and $B \subseteq A$. When this happens we, unsurprisingly, write $A = B$.

**Example 1.3.**

Let $A$ and $B$ be the sets described below:

$$A = \left\{x \in \mathbb{Z} \,\middle|\, x = 2n \text{ for some } n \in \mathbb{Z}, \text{ and } x = 3m \text{ for some } m \in \mathbb{Z}\right\},$$
$$B = \left\{y \in \mathbb{Z} \,\middle|\, y = 6n \text{ for some } n \in \mathbb{Z}\right\}.$$

Show that $A$ and $B$ are equal.

Here, $A$ is the set of all integers which are simultaneously multiples of 2 and 3, while $B$ is the set of all integers which are multiples of 6. If you start writing down a few elements of $A$, then you'll probably be convinced pretty quickly that, sure enough, everything in $A$ is a multiple of 6, but let's actually prove this.

We first want to show that $A \subseteq B$: i.e., every integer which is a multiple of both 2 and 3 must be a multiple of 6. So suppose $x \in A$, we want to show that $x \in B$ as well. If $x \in A$ then $x = 2n = 3m$ for some pair of integers $m$ and $n$. This equation means, in particular, that 2 divides $3m$. Since 2 is a prime number it must divide either 3 or $m$ (this is basically the definition of a prime number; see the Wikipedia page about prime numbers for more information). Since 2 does not divide 3, it must divide $m$. Thus $m = 2k$ for some $k$. This means $x = 3m = 3 \cdot 2k = 6k$, and so $x$ must be a multiple of 6. Hence if $x \in A$, then $x \in B$ as well, so $A \subseteq B$.

We also need to show that $B \subseteq A$. Suppose that $y \in B$, so $y = 6k$ for some $k$. But then $y = 3 \cdot 2 \cdot k$, and so $y$ is simultaneously a multiple of 2 (take $n = 3k$ in the definition of $A$) and a multiple of 3 (let $m = 2k$). Thus $B \subseteq A$.

As $A \subseteq B$ and $B \subseteq A$, $A = B$.

**Remark.**
Just a reminder that it's okay if you don't understand an example when you first read it in these notes. The important thing is to make an effort to try to understand it. Usually just making an effort, even if you don't feel comfortable that you understood what you just read, still helps to get your brain thinking about the idea. You may find that if you read something you don't understand, then step away from it for a while (a few hours, maybe a day or two) and then re-read it, it might make sense on the second reading. If you still don't understand

the example on a second reading, don't beat yourself up about it. Feel free to ask questions about the idea through email, office hours, or in class if you're still unable to understand what's going on. The most important thing is to keep trying and not let one thing you don't understand discourage you from trying anything else.

## 1.5 The empty set

There is one special set in mathematics called ***the empty set*** which is the only set that contains no elements; it is the only set of cardinality zero and is denoted $\emptyset$.

A set without anything in it might sound uninteresting, but there is at least one surprising thing about the empty set: the empty set is a subset of every other set. That is, for any set $A$, $\emptyset \subseteq A$. Why is this the case? We should only write $\emptyset \subseteq A$ if every element of $\emptyset$ is also an element of $A$. Since $\emptyset$ has no elements, however, it immediately satisfies this definition! All the elements of $\emptyset$ (all zero of them) are also elements of $A$!

**Exercise 1.4.**
If the idea that the empty set is a subset of every other set sounds a little bit odd, re-read the above paragraph and think about the logic behind the last sentence until it makes sense. The solution to this exercise in the appendices gives another way to think about this if the first explanation above simply won't "click" for you.

## 1.6 Real numbers

So far we have described three different sets of numbers: the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, and the rational numbers $\mathbb{Q}$. We now describe one more set of numbers which we will use in this class: the real numbers.

To define the real numbers rigorously would take us very far afield, and so we will be a little bit hand-wavy in the definition. A ***real number*** is simply the coordinate of a point on the real line; equivalently, it is the collection of all numbers that we can write down with a (possibly infinite)

decimal expansion. All of the numbers described thus far (natural numbers, integers, and rational numbers) are real numbers: we can write 6 as 6.000...; we can write -3 as -3.000...; we can write $\frac{22}{7}$ as 3.142857142857142857...

The set of all real numbers is denoted $\mathbb{R}$. Notice we have the following string of subsets:

$$\mathbb{N} \subsetneq \mathbb{Z} \subsetneq \mathbb{Q} \subsetneq \mathbb{R}.$$

Notice that the examples of real numbers we wrote down above all have a decimal expansion which is eventually repeating. However there are numbers that can't be written in this way. One simple example is $\sqrt{2}$. We can write $\sqrt{2}$ as an infinite decimal expansion $\sqrt{2} = 1.414213562...$, this expansion never repeats no matter how many digits you write down. We won't do it, but it can be shown that every rational number has an eventually repeating decimal expansion. So, another way to say that the decimal expansion of $\sqrt{2}$ never repeats, is to say that it is impossible to write $\sqrt{2}$ as a ratio of two integers. That is, $\sqrt{2}$ is not a rational number. A real number that is not rational is called an ***irrational number***.

**Theorem 1.1.**
*The square root of* 2 *is irrational.*

We will prove Theorem 1.1 using a a standard proof technique technique called ***proof by contradiction***, which is also sometimes called *reductio ad absurdum*, Latin for "reduction to absurdity." The idea is that we will suppose that what we want to show to be true is in fact false, and show that this leads to some impossible situation. Since if the statement were false something impossible would have to happen, it must be that the statement is in fact true.

**Remark.**
You really don't need to try to understand the proof below if you don't want to. It is included only for the sake of completeness and to justify the claim that not all real numbers are rational. If you're willing to take this on faith, you can safely skip over the proof below.

*Proof of Theorem 1.1.*
Suppose that $\sqrt{2}$ were rational. Then we could write $\sqrt{2} = \frac{p}{q}$ where $p$ and $q$ were integers. We could cancel out any common factors of $p$ and $q$, and so we may assume that there is nothing which divides both $p$ and $q$ simultaneously.

If we square both sides of the equation $\sqrt{2} = \frac{p}{q}$, then we obtain $2 = \frac{p^2}{q^2}$. This means $q^2 = \frac{p^2}{2}$. As $q^2$ is an integer, $\frac{p^2}{2}$ must be an integer as well. This means that $p^2$ must be even. If $p^2$ is even, however, then $p$ must also be even.[a] Thus $p = 2r$ for some integer $r$. But then $p^2 = 4r^2$, so $q^2 = \frac{p^2}{2} = 2r^2$. Again, $q^2$ is even so $q$ must also be even: say $q = 2s$.

At this point we have a contradiction: we originally assumed that $p$ and $q$ had no common factors, but then showed that $p$ and $q$ must both be even meaning that 2 is a common factor. This is impossible: $p$ and $q$ can not simultaneously have no common factors and also have 2 as a common factor! We obtained this contradiction because we supposed that $\sqrt{2}$ was rational, so we must conclude that $\sqrt{2}$ is not rational. □

[a] If this isn't clear, think about it and notice that 2 is a prime number.

The main takeaway from this is that there are numbers we care about, the real numbers, which are not rational numbers. The vast majority of the time this won't really matter for our purposes in this class, but it's good to be aware of this fact.

## 1.7 Practice problems

**Problem 1.1.**
Rewrite each set below using set-builder notation.

(a) $\{1, 4, 9, 16, 25, 36, 49, 64, 81, 100, ...\}$

(b) $\{-1, 4, -9, 16, -25, 36, -49, 64, -81, 100, ...\}$

(c) The set of rational numbers where the numerator is the cube of the denominator.

(d) The set of points in the plane which are on the graph of the function $f(x) = x^3$.

**Problem 1.2.**
Let $A$ be the set of all natural numbers which are multiples of 15, $B$ be the set of all natural numbers which are multiples of 10, $C$ the set of all natural numbers which are multiples of 20, and $D$ the set of all natural numbers which are multiples of 30.

(a) Write $A$, $B$, $C$, and $D$ in set builder notation.

(b) Is $A \subseteq B$? Explain why or why not.

(c) Is $A \subseteq C$? Explain why or why not.

(d) Is $A \subseteq D$? Explain why or why not.

(e) Is $B \subseteq A$? Explain why or why not.

(f) Is $B \subseteq C$? Explain why or why not.

(g) Is $B \subseteq D$? Explain why or why not.

(h) Is $C \subseteq A$? Explain why or why not.

(i) Is $C \subseteq B$? Explain why or why not.

(j) Is $C \subseteq D$? Explain why or why not.

(k) Is $D \subseteq A$? Explain why or why not.

(l) Is $D \subseteq B$? Explain why or why not.

(m) Is $D \subseteq C$? Explain why or why not.

**Problem 1.3.**
Let $A$ be the set of all points $(x, y)$ in the circle of radius one centered at the origin, and let $B$ be the set of all points $(x, y)$ satisfying the inequality $x^2 + \frac{y^2}{4} \leq 1$. Show $A \subseteq B$.

**Problem 1.4.**
Are there any sets $A$ such that $A \subseteq \emptyset$? If not, why not? If so, what can be said about such a set?

**Problem 1.5.**
Let $A$ be the graph of the function $f(x) = \frac{x^2-1}{x-1}$, and $B$ the graph of the function $g(x) = x + 1$. Are $A$ and $B$ the same sets? If not, is one a subset of the other?

**Problem 1.6.**
Suppose $A \neq B$. Is it true that there must be an element of $A$ which is not an element of $B$, *and* an element of $B$ which is not an element of $A$?

# 2 Operations on Sets

*"Contrariwise," continued Tweedledee, "if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."*

LEWIS CARROLL
*Through the Looking Glass*

## 2.1 Unions

Given a collection of sets there are many different ways we can combine the sets together to get new sets. Here we discuss the three most important such operations: unions, intersections, and products.

Given two sets $A$ and $B$, their ***union*** is the "smallest" set which contains every element of $A$ as well as every element of $B$, and is denoted $A \cup B$. You should think of the union as gluing two sets together to get a bigger set.

**Example 2.1.**
Let $A = \{2, 4, 6, 8, 10\}$ be the set of all even integers between 1 and 10, and let $B = \{1, 3, 5, 7, 9\}$ be the set of all odd integers between 1 and 10. Then their union $A \cup B$ is the set of all integers between 1 and 10:

$$A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

**Exercise 2.1.**
Let $A$ and $B$ be any two sets. Show that $A \subseteq A \cup B$ and $B \subseteq A \cup B$.

Given any two sets $A$ and $B$, there are going to be *lots* of other sets that contain $A$ and $B$ as subsets. In the example above, for instance, the set

$$\{1, 2, ..., 10, 11\}$$

contains both $A$ and $B$ as a subset, as does $\mathbb{N}$ and $\mathbb{Z}$. The union $A \cup B$ is the *smallest* set containing both $A$ and $B$ as subsets in the following sense: If $A \subseteq C$ and $B \subseteq C$, then $A \cup B \subseteq C$.

## 2.2 Intersections

Another operation we can perform on two sets is to intersect them. The ***intersection*** of two sets $A$ and $B$, denoted $A \cap B$, consists precisely of all of the elements which are in both $A$ and $B$. That is, $x \in A \cap B$ if and only if $x \in A$ and $x \in B$.

**Example 2.2.**
Let $A$ be the set of all multiples of 6, and $B$ the set of all multiples of 10,

$$A = \{... - 18, -12, -6, 0, 6, 12, 18, ...\},$$
$$B = \{... - 30, -20, -10, 0, 10, 20, 30, ...\}.$$

Then $A \cap B$ is the set of all the numbers which are both multiples of 6 and 10.

$$A \cap B = \{..., -90, -60, -30, 0, 30, 60, 90, ...\}.$$

**Example 2.3.**
Suppose that $S$ is the set of all characters that have ever appeared in a Star Wars film,

$$S = \{\text{Luke Skywalker, Obi-Wan Kenobi, Kylo Ren, } \cdots\},$$

that $R$ is the set of all droids from the Star Wars films,

$$R = \{\text{R2D2, C3P0, BB-8,} \cdots\},$$

$D$ is the set of all characters corrupted by the dark side of The Force,

$$D = \{\text{Darth Vader, Kylo Ren, Emperor Palpadine,} ...\}$$

and $V$ is the set of all characters which appeared in *Star Wars V: The Empire Strikes Back*,

$$V = \{\text{Luke Skywalker, Lando Calrissian, Boba Fett}, \cdots\}.$$

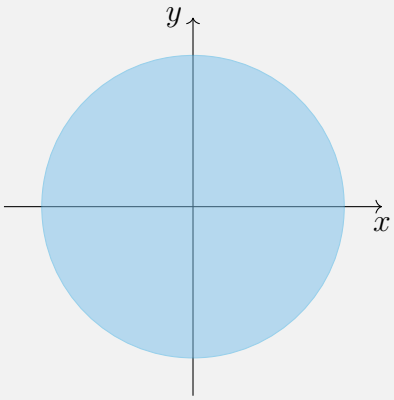
Then the set of all droids that appeared in *The Empire Strikes Back* is the intersection of the set of all droids and the set of all characters that were in that movie:

$$R \cap V = \{\text{C3P0, R2D2}\}.$$

The set of all characters which were corrupted by the dark side of The Force and were in *The Empire Strikes Back* is the intersection of all characters corrupted by the dark side of The Force and the set of all characters in *The Empire Strikes Back*:

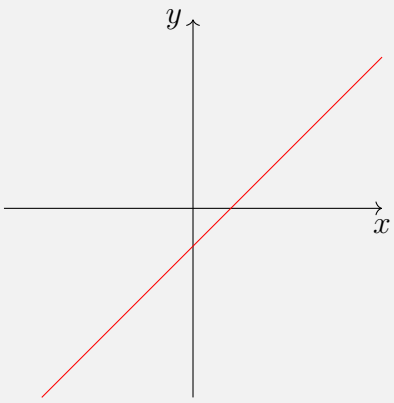
$$D \cap V = \{\text{Darth Vader, Emperor Palpadine}\}.$$

**Example 2.4.**
Suppose that $A$ is the set of all points in the plane (all $(x, y)$-pairs) that are at most distance 1 from the origin.



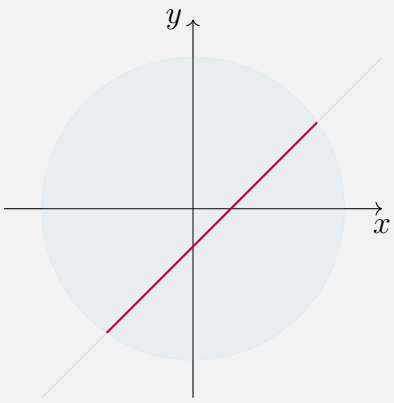

$$A = \left\{(x, y) \,\middle|\, x^2 + y^2 \leq 1\right\}$$

And suppose that $B$ is the line with slope 1 through the point $(0, -0.25)$




$$B = \{(x, y) \mid y = x - 0.25\}$$

Then the intersection $A \cap B$ is then the portion of the line $B$ that remains inside the disc $A$. This is the dark purple line segment in the figure below. (The original disc and line are drawn in very lightly just for comparison; they are not part of $A \cap B$.)




$$A \cap B = \{(x, y) \mid y = x - 0.25 \text{ and } x^2 + y^2 \leq 1\}$$

It may happen that two sets have nothing in common: for example, the set $A = \{1, 2, 3\}$ and the set $B = \{4, 5, 6\}$ have no common elements. In a situation such as the intersection of the two sets is empty, $A \cap B = \emptyset$, and we say that $A$ and $B$ are ***disjoint***.

**Exercise 2.2.**
Let $A$ and $B$ be any two sets. Show that $A \cap B$ is a subset of $A$ and also a subset of $B$.

**Exercise 2.3.**
Show that if $A \subseteq B$, then $A \cap B = A$.

Just as the union $A \cup B$ was the smallest set containing both $A$ and $B$ as subsets, the intersection $A \cap B$ is the largest subset of both $A$ and $B$ in the following sense: If $C \subset A$ and $C \subseteq B$, then $C \subseteq A \cap B$.

Anytime you have several operations defined on some collection of objects (e.g., unions and intersections defined for sets), you might be interested in how those operations interact with one another. For unions and intersections this interaction is similar distributive law for normal numbers (e.g., that $x \cdot (y + z) = x \cdot y + x \cdot z$).

**Proposition 2.1.**
*For any sets $A$, $B$, and $C$ we have the following two distributive laws:*

$$A \cap (B \cup C) = [A \cap B] \cup [A \cap C]$$
$$A \cup (B \cap C) = [A \cup B] \cap [A \cup C]$$

*Proof.*
We will only prove the first distributive law; the proof of the second one is almost identical.

Notice that elements of $A \cap (B \cup C)$ are elements of $A$ which are also elements of either $B$ or $C$. The elements of $A \cap B$ are elements of both $A$ and $B$; the elements of $A \cap C$ are elements of both $A$ and $C$. Unioning $A \cap B$ and $A \cap C$ together, we have exactly the elements of

$A$ which are also in either $B$ or $C$. □

## 2.3 Products

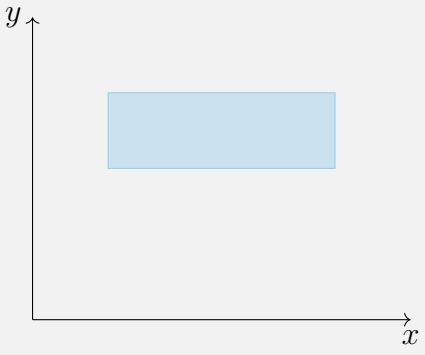
One last operation we will mention is the Cartesian product, which we will usually refer to simply as the "product." Given two sets, $A$ and $B$, their ***(Cartesian) product*** is a set denoted $A \times B$ and which consists of all ordered pairs $(a, b)$ where $a \in A$ and $b \in B$:

$$A \times B = \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

**Example 2.5.**
Let $A = \{x, y, z\}$ and $B = \{u, v, x\}$. Then

$$\begin{aligned} A \times B = \{&(x, u),\ (x, v),\ (x, x),\\ &(y, u),\ (y, v),\ (y, x),\\ &(z, u),\ (z, v),\ (z, x)\} \end{aligned}$$

**Example 2.6.**
Let $A$ be the interval $[1, 4]$ and $B$ the interval $[2, 3]$. Then the product $A \times B$ consists of all pairs of numbers (i.e., all $(x, y)$ pairs in the plane) where the first coordinate is between 1 and 4, and the second coordinate is between 2 and 3:



$$A \times B = \big\{(x, y) \,\big|\, 1 \leq x \leq 4 \text{ and } 2 \leq y \leq 3\big\}$$

It is fairly often that we will want to consider the product of a set with itself, $A \times A$. In such a situation we will usually simply write $A^2$ to mean $A \times A$.

The three operations we described above can be defined for more than two sets. For example, it makes sense to talk about the union, intersection, or product of three sets. It is completely reasonable, for example, to say that the union $A \cup B \cup C$ should be the smallest set containing all the elements of $A$, all the elements of $B$, and all the elements of $C$. The intersection $A \cap B \cap C$ should contain only those elements that are in all three sets $A$, $B$, and $C$.

**Example 2.7.**
Consider the sets $A$, $B$, and $C$ described below:

$$A = \{1, 2, 3, ..., 10\}$$
$$B = \{2, 4, 6, ..., 20\}$$
$$C = \{-12, -9, -6, ..., 6, 9, 12\}.$$

The union of these sets is

$$A \cup B \cup C = \{ -12, -9, -6, -3, 0, 1, 2, 3, ...10, \\ 12, 14, 16, 18, 20\}.$$

The intersection is

$$A \cap B \cap C = \{6\}.$$

Of course, there's nothing magical about having two sets or three sets: we can define unions and intersections for any number of sets – even infinitely-many.

**Example 2.8.**
For each $n \in \mathbb{N}$ define the set $A_n$ to be the interval $\left[-\frac{1}{2^n}, \frac{1}{2^n}\right]$. The first few intervals are thus

$$\begin{aligned} A_1 &= [-1/2,\ 1/2] \\ A_2 &= [-1/4,\ 1/4] \\ A_3 &= [-1/8,\ 1/8] \\ A_4 &= [-1/16,\ 1/16] \\ &\vdots \end{aligned}$$

The intersection of all these intervals is usually written in one of two ways,

$$A_1 \cap A_2 \cap A_3 \cap \cdots \quad \text{or} \quad \bigcap_{n=1}^{\infty} A_n,$$

and consists of all the elements which are in *every* $A_n$. In this case the only such element is 0:

$$\bigcap_{n=1}^{\infty} A_n = \{0\}\,.$$

**Exercise 2.4.**
For each $n \in \mathbb{N}$, let $B_n$ be the following interval:

$$B_n = \left[\frac{1}{2^n},\ 1 - \frac{1}{2^n}\right].$$

What is the infinite union of all the $B_n$'s, $\bigcup_{n=1}^{\infty} B_n$?

The product might be slightly, but not very, surprising. When we write a product of three sets we will mean the collection of ordered triples; a product of four sets is the collection of ordered quadruples. In general, the product of $n$ sets is the set of all ordered $n$-tuples. (An ***n-tuple*** is an ordered list of $n$ items. A 2-tuple is simply a pair; a 3-tuples is a triple; a 5-tuple has the form $(a, b, c, d, e)$.)

**Example 2.9.**
Let $A$, $B$, and $C$ be the following sets:

$$A = \{1, 2, 3\}$$
$$B = \{\alpha, \beta\}$$
$$C = \{\sharp, \flat\}$$

Then $A \times B \times C$ is the following set

$$\{(1, \alpha, \sharp), (1, \alpha, \flat), (1, \beta, \sharp), (1, \beta, \flat),$$
$$(2, \alpha, \sharp), (2, \alpha, \flat), (2, \beta, \sharp), (2, \beta, \flat),$$
$$(3, \alpha, \sharp), (3, \alpha, \flat), (3, \beta, \sharp), (3, \beta, \flat)\}$$

It is very common to consider the Cartesian product of a set $A$ with itself $n$ times, so we usually denote this as $A^n$.

**Example 2.10.**
The set of all ordered triples of integers could be written $\mathbb{Z}^3$:

$$\mathbb{Z}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{Z}\}$$

We can also talk about products of infinitely-many sets, but for simplicity we will avoid that for the time being.

## 2.4 Complements

The last operation on sets we will describe is not defined for all sets, but only for subsets of some given set. That is, in some applications there will be some ambient set "in the background," and all other sets we are interested in will be subsets of this ambient set. In such a situation, we sometimes call the ambient set the ***universe*** because it consists of everything we care about for the problem at hand. For example, in geometry the universe may be the set of all points in the plane, $\mathbb{R}^2$ – for some geometric problems everything you care about might take place in the plane, so that is your universe.

Once we have a universal set $\mathcal{U}$, we can define the *complement* of any subset $E \subseteq \mathcal{U}$, which you should think of as being the complete opposite of $E$. To be more precise, given any set $E$ inside the universe $\mathcal{U}$, the complement of $E$, denoted $E^c$, is the set of all elements in $\mathcal{U}$ which are not in $E$:

$$E^c = \left\{ x \in \mathcal{U} \,\middle|\, x \notin E \right\}.$$

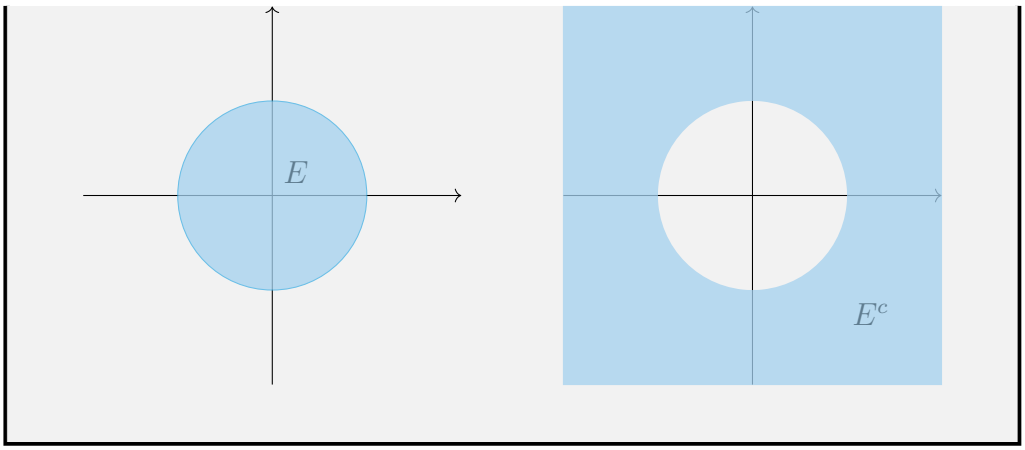
**Example 2.11.**
Suppose the universe $\mathcal{U}$ consists of all integers between 1 and 10,

$$\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

- If $E$ is the set of all even numbers between 1 and 10, $E = \{2, 4, 6, 10\}$, then its complement consists of all the odd numbers, $E^c = \{1, 3, 5, 7, 9\}$.
- if $E$ is the set of all numbers in $\mathcal{U}$ greater than 7, $E = \{8, 9, 10\}$, then its complement is the set of all numbers less-than-or-equal-to 7, $E^c = \{1, 2, 3, 4, 5, 6, 7\}$.

**Exercise 2.5.**
Let $\mathcal{U}$ be any universal set and $E \subseteq \mathcal{U}$ any subset. Show $(E^c)^c = E$.

**Example 2.12.**
Suppose the universe $\mathcal{U}$ consists of all points in the plane, $\mathcal{U} = \mathbb{R}^2$. If $E$ is the set of all points whose distance to the origin is at most 1 (so, $E$ is the circular disc of radius 1 centered at the origin), then its complement $E^c$ consists of all the points distance more than 1 from the origin (this would be the entire plane with a "hole" of radius 1 centered at the origin).



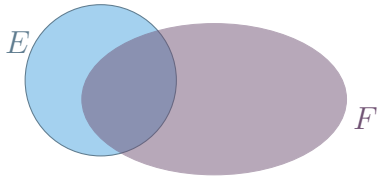

**Exercise 2.6.**
Given some universe $\mathcal{U}$, what is the complement of the empty set $\emptyset$? What is the complement of $\mathcal{U}$?

## 2.5 Difference

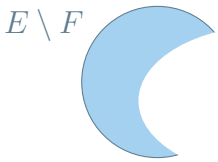
The difference between two sets $E$ and $F$, denoted $E \setminus F$, is the set of all elements in $E$ which are not also elements in $F$:

$$E \setminus F = \left\{ x \in E \,\middle|\, x \notin F \right\}.$$

To have a picture of this, imagine that $E$ and $F$ are the overlapping regions indicated below.




Then the set difference $E \setminus F$, is the shaded region below.




**Example 2.13.**
Let $S$ be the set of all Star Wars movies,

$$S = \{\text{Star Wars}, \text{The Empire Strikes Back}, \text{Return of the Jedi}, \\ \text{The Phantom Menace}, \text{The Clone Wars}, \text{Revenge of the Sith}, \\ \text{The Force Awakens}, \text{Rogue One}, \text{The Last Jedi}, \\ \text{Solo}\}$$

and let $D$ be the set of all movies produced by Disney,

$$D = \{\text{Snow White}, \text{Pinocchio}, ..., \text{Coco}, \text{The Force Awakens}, ...\}.$$

Then $S \setminus D$ would be the set of all Star Wars movies not produced by Disney,

$$S \setminus D = \{\text{Star Wars}, \text{The Empire Strikes Back}, \text{Return of the Jedi}, \\ \text{The Phantom Menace}, \text{The Clone Wars}, \text{Revenge of the Sith}\}$$

**Exercise 2.7.**
Show that $E \setminus F$ is equal to $E \setminus (F \cap E)$.

## 2.6 De Morgan's laws

It is very common in mathematics to have multiple possible operations you can perform on a given type of object, and then to ask how these operations interact with one another. For example, in arithmetic two basic operations

are addition and multiplication, and these two operations "interact" via the distributive law $a \cdot (b + c) = a \cdot b + a \cdot c$.

At this point we have a few different operations we can perform on sets, and we want to know how they interact with each other. In particular, we have unions, intersections, and complements. These three operations are related by two rules called ***de Morgan's laws***, which essentially say that unions turn into intersections (and intersections turn into unions) when we take complements.

More precisely, if $E$ and $F$ are two subsets of some universe $\mathcal{U}$ (recall we always need a "universe" to discuss complements), then we have the following:

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

That is, we can intentionally turn unions into intersections and vice versa, but we also have to take the complement of the sets involved. Right now it might be hard to appreciate why this is something we'd like to do, but we'll see later that when calculating probabilities we will have special rules for calculating probabilities of unions and intersections. In some types of problems we use de Morgan's laws to turn a complicated problem involving probabilities of unions into a simpler problem involving probabilities of intersections. (This is a little ways down the road from where we are now, but that's where we're heading.)

**Example 2.14.**
Suppose the universal set is $\mathcal{U} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and let $E = \{1, 2, 3\}$ and $F = \{3, 4, 5\}$. Verify directly that de Morgan's laws are satisfied.

Here we just want to compute the four sets stated in de Morgan's laws above (two sets per equation) and see if the equalities that are claimed to be true are in fact satisfied.

First note $E \cup F = \{1, 2, 3, 4, 5\}$. Hence $(E \cup F)^c = \{6, 7, 8, 9, 10\}$. Now note $E^c = \{4, 5, 6, 7, 8, 9, 10\}$ and $F^c = \{1, 2, 6, 7, 8, 9, 10\}$. Their intersection is $E^c \cap F^c = \{6, 7, 8, 9, 10\}$. So the first equation in de Morgan's laws is satisfied.

For the second equation we note $E \cap F = \{3\}$, so $(E \cap F)^c = \{1, 2, 4, 5, 6, 7, 8, 9, 10\}$. Now $E^c$ and $F^c$ we computed above, and their

union is $E^c \cup F^c = \{1, 2, 4, 5, 6, 7, 8, 9, 10\}$, and so the second equation in de Morgan's laws is satisfied.

The proof of de Morgan's laws essentially has to do with working out what each side of each equation means. We will simply prove the first law, leaving the second one as an exercise.

*Proof of de Morgan's first law.*
We wish to show that $(E \cup F)^c = E^c \cap F^c$. To show two sets are equal we must show each one is a subset of the other: i.e., we must show $(E \cup F)^c \subseteq E^c \cap F^c$ and also that $E^c \cap F^c \subseteq (E \cup F)^c$.

Let $x \in (E \cup F)^c$. That is, $x$ is an element of $\mathcal{U}$ which is in neither $E$ nor $F$. Since $x \notin E$ and $x \notin F$, we have $x \in E^c$ and $x \in F^c$, so $x \in E^c \cap F^c$. This shows $(E \cup F)^c \subseteq E^c \cap F^c$.

Now to show the other inclusion, let $x \in E^c \cap F^c$. Thus $x$ is in both $E^c$ and $x$ is in $F^c$. This means $x$ is in neither $E$ nor $F$, and hence $x \notin E \cup F$. By the definition of the complement, that means $x \in (E \cup F)^c$. Hence $E^c \cap F^c \subseteq (E \cup F)^c$. □

**Exercise 2.8.**
Prove the second law of de Morgan. That is, if $E$ and $F$ are subsets of a universal set $\mathcal{U}$, then $(E \cap F)^c = E^c \cup F^c$.

There's nothing really special about our using two sets in the statements of de Morgan's laws above instead of three or four or five or ... In general, given any collection of subsets $E_1$, $E_2$, ..., $E_n$ of some universal set $\mathcal{U}$, de Morgan's laws extend to

$$(E_1 \cup E_2 \cup \cdots \cup E_n)^c = E_1^c \cap E_2^c \cap \cdots \cap E_n^c$$
$$(E_1 \cap E_2 \cap \cdots \cap E_n)^c = E_1^c \cup E_2^c \cup \cdots \cup E_n^c.$$

If you believe the proof of de Morgan's laws for two sets, then it's easy to see how to get de Morgan's laws for more than two sets. Let's consider

the case when there are three sets, and let's just call them $E$, $F$, and $G$. The first law says that

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

How can we get this if we know only have de Morgan's laws for two sets? We'll just cheat and rewrite the above as two sets. If we write $H = E \cup F$, then $E \cup F \cup G$ can be written as $H \cup G$. De Morgan's laws on two sets then tell us

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

Now let's figure out what $H^c$ is: since $H = E \cup F$, we must have $H^c = (E \cup F)^c$. But now de Morgan's laws for two sets tell us $H^c = (E \cup F)^c = E^c \cap F^c$. Plugging this in for $H^c$ on the right-hand side above we have

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

The same trick works for de Morgan's second law for three sets.

Now that we know de Morgan's laws for three sets, it's easy to extend it to de Morgan's law for four sets; once we have de Morgan's laws for four sets, we can easily extend to five sets; etc. We just keep taking a "complicated" de Morgan's law with lots of sets and rewriting it in terms of de Morgan's law with fewer sets. Repeating this process several times we can always boil everything back down to de Morgan's law with two sets which we already know.

**Remark.**
This idea of taking a complicated problem and breaking it up into slightly simpler problems which you can continue to break up into slightly simpler problems until you get down to the simplest possible scenario is very common in mathematics and in computer science. In math this is usually called ***induction***, whereas in computer science it's called ***recursion***, but they're really the same thing.

## 2.7 Practice problems

**Problem 2.1.**
Let $E$ be the set of all even integers, and $F$ the set of all integer multiples of five. Find a simple way to express $E \cap F$ in set-builder notation.

**Problem 2.2.**
Suppose $A$ and $B$ are sets that satisfy the following: $A \subseteq A \cap B$. What does this tell you about $B$?

**Problem 2.3.**
Determine the following infinite intersection,

$$\bigcap_{n=1}^{\infty} (-\infty, -n)$$

**Problem 2.4.**
For each positive real number $a$, let $H_a$ denote the set of points in the $xy$-plane whose $y$-coordinate is greater than or equal to $a$,

$$H_a = \left\{ (x, y) \,\middle|\, y \geq a \right\}.$$

What is

$$\bigcup_{a>0} H_a?$$

# 3

# *Functions*

*When you have eliminated the impossible, whatever remains, however improbable, must be the truth.*

---

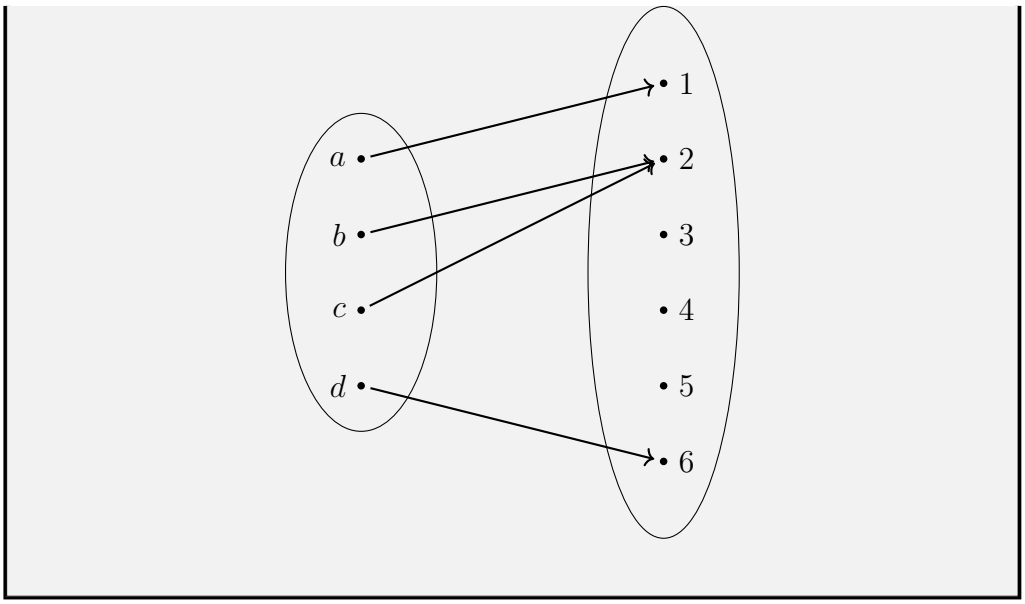
Sir Arthur Conan Doyle
*The Sign of Four*

The material in this chapter will be crucial when we begin to discuss random variables later, however we will not need this material until then. You may want to only skim over this chapter for now, then return to it once we begin discussing random variables in class.

## 3.1 Definitions and examples

Given two sets $A$ and $B$, a ***function*** from $A$ to $B$ (also called a ***map*** from $A$ to $B$) is a rule which associates to each element of $A$ an element of $B$. Sometimes we will represent functions pictorially by drawing $A$ on the left, $B$ on the right, and then having arrows going from elements of $A$ to elements of $B$.

**Example 3.1.**
Suppose $A = \{a, b, c, d\}$ and $B = \{1, 2, ..., 6\}$. One possible function from $A$ to $B$ is pictured below:




The function in Example 3.1 associates 1 to $a$; associates 2 to $b$; 2 is also associated to $c$; and finally $d$ gets associated to 6.

It is convenient to give a function a name so that we can refer to it without drawing pictures like this all of the time. Let's refer to the function from Example 3.1 as $f$. To say that $f$ takes elements of $A$ and associates an element of $B$ to them we write $f : A \to B$. We then call $A$ the ***domain*** of $f$ and $B$ is called the ***codomain*** of $f$. The ***range*** of $f$ is the subset of $B$ which actually get associated to an element of $A$. For the function in Example 3.1 the range is $\{1, 2, 6\}$.

There are several different notations that are used to describe which elements of $B$ a function associates to elements of $A$. Some commonly used ones are $f(a) = b$ and $a \mapsto b$. The first one you've probably seen before, but the second one might be new. We pronounce $a \mapsto b$ as "$a$ maps to $b$."

**Example 3.2.**
Considering the function $f$ shown in Example 3.1 we have

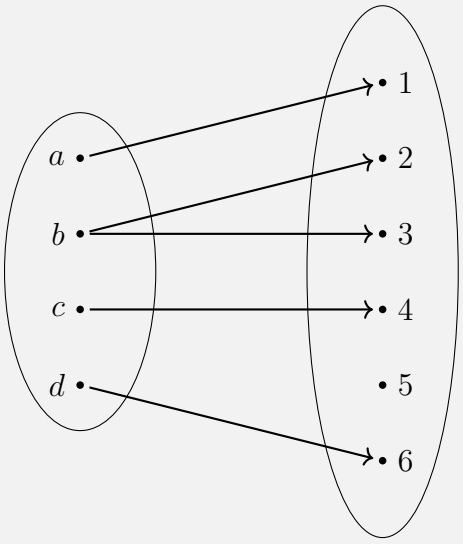
$$f(a) = 1$$
$$f(b) = 2$$
$$f(c) = 2$$
$$f(d) = 6$$

Using the other notation we would write

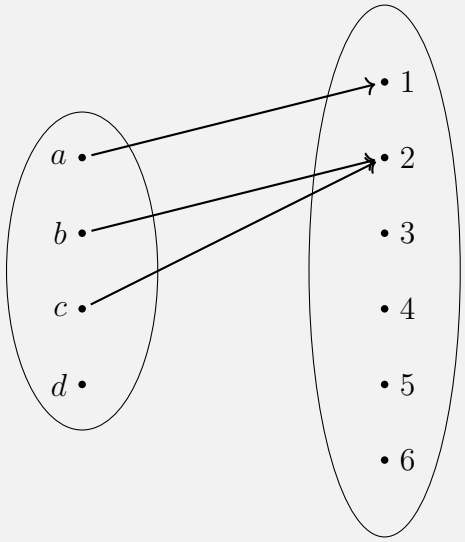
$$a \mapsto 1$$
$$b \mapsto 2$$
$$c \mapsto 2$$
$$d \mapsto 6$$

It is important to realize that a function $f : A \to B$ can only associate one element of $b$ to a given element of $a$ (even though there could be several elements of $A$ associated to a given $b \in B$). A function $f : A \to B$ must also associate *every* element of $A$ to something in $B$, even though not every element of may $B$ necessarily have something associated to it. (The range of $f : A \to B$ is by definition the set of all elements in $B$ which have an element of $A$ associated to them.)

**Example 3.3.**
The following is not a function:



**Example 3.4.**
The following is not a function:




## 3.2 Representing functions

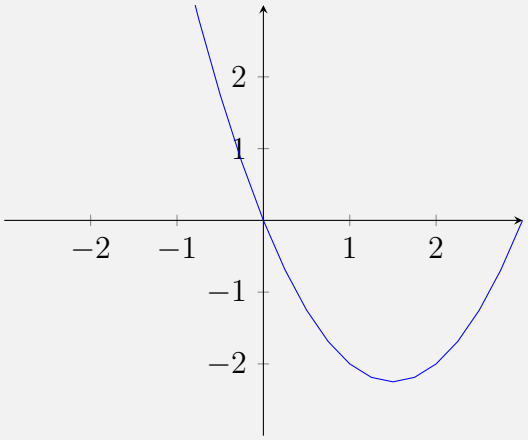
It is common to represent a function by a formula, for example consider the function $f : \mathbb{Z} \to \mathbb{Z}$ which takes a given number and squares it. It's not really reasonable to represent this function pictorially since $\mathbb{Z}$ has infinitely-many elements, so we instead describe the function by an algebraic rule and write $f(x) = x^2$ or $x \mapsto x^2$.

Another way to represent a function is to consider its graph. In general, the ***graph*** of a function $f : A \to B$, which we will denote $\text{Graph}(f)$, is a subset of $A \times B$ which consists of pairs of the form $(a, f(a))$. That is,

$$\text{Graph}(f) = \left\{(a, b) \in A \times B \,\middle|\, b = f(a)\right\}.$$

When we have a function from the set of real numbers $\mathbb{R}$ (defined below) to itself, it is common to actually draw these points in the plane $\mathbb{R}^2$. That is, given a function $f(x)$ we plot all of the pairs $(x, y)$ where $y = f(x)$.

**Example 3.5.**
The graph of the function $f : \mathbb{R} \to \mathbb{R}$ given by $x \mapsto x^2 - 3x$ is

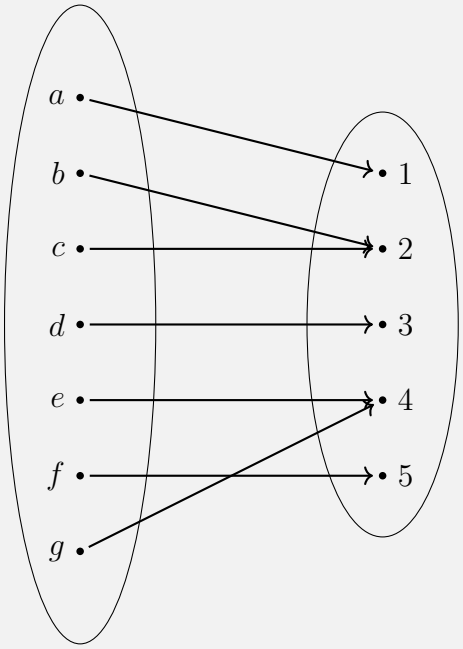



## 3.3 Special types of functions

As mentioned above, a function $f : A \to B$ must associate every element of $A$ to some element of $B$ (i.e., for every $a \in A$, $f(a)$ is defined), but not every element of $B$ must have an element of $A$ associated to it (there may be some $b \in B$ such that for every $a \in A$, $f(a) \neq b$). In the special case where every element of $B$ *does* have an element of $a$ associated to it, we say the map $f$ is ***surjective*** or ***onto***. Equivalently, a function is surjective when its codomain and range are the same.

**Example 3.6.**
The function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(x) = x + 3$ is surjective. Every $y$ in the codomain $\mathbb{Z}$ gets associated an $x$ from the domain, namely $x = y - 3$.

**Example 3.7.**
The function pictured below is surjective.

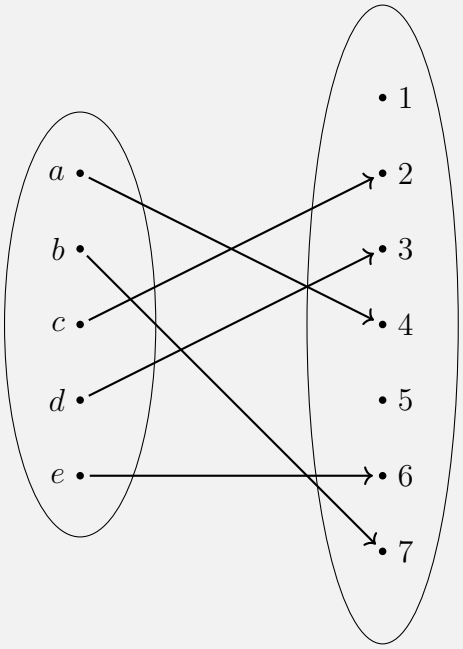



**Remark.**
The terms *surjective* and *onto* are completely synonymous, and which one a person uses is largely a matter of personal preference.

Notice that the function pictured in Example 3.7 has the property that multiple elements of the domain get associated to the same element in the codomain: both $b$ and $c$ get associated to 2, while both $e$ and $g$ are associated to 4. When this *does not* happen, we give the function a special name.

We say that a function $f : A \to B$ is ***injective*** or ***one-to-one*** (commonly denoted ***1-1***) if each element of $A$ is associated to a unique of $B$. That is, if $a_1$ and $a_2$ are distinct elements of $A$, then $f(a_1) \neq f(a_2)$.

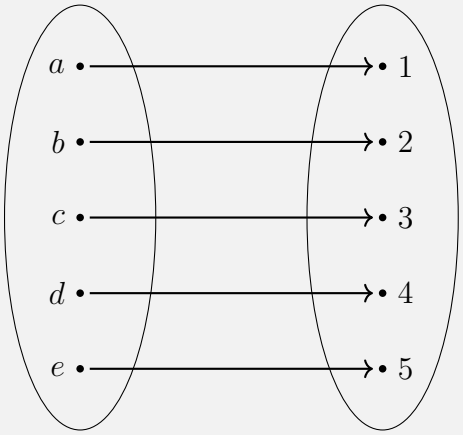
**Example 3.8.**
The following function is injective.




**Example 3.9.**
The function $f(x) = x + 3$ from Example 3.6 is injective: if $x_1 \neq x_2$, then $f(x_1) = x_1 + 3 \neq x_2 + 3 = f(x_2)$.

When a function is both injective and surjective, we say the function is ***bijective***. Bijective functions play a special role in most areas of mathematics because having a bijection between two sets means those two sets are "the same." That is, you may label the elements of the sets differently and think of them in different ways, but each element in one set has exactly one element in the other set associated to it: we can pair the elements of the sets together one by one.

**Example 3.10.**
The following function is injective.




**Example 3.11.**
The function $f(x) = x + 3$ from Example 3.6 is bijective as it is both surjective and injective.

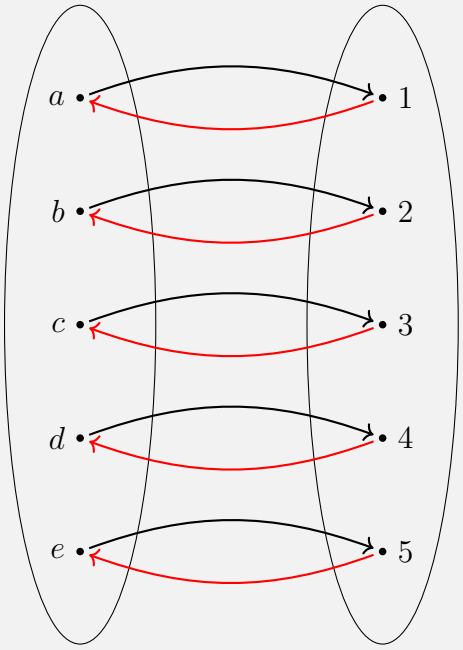
**Remark.**
If we know that a given function $f : A \to B$ is injective, surjective, or bijective, then we also instantly know how the cardinalities of $A$ and $B$ are related. If $f$ is injective, then $\#A \leq \#B$. If $f$ is surjective, then $\#A \geq \#B$. If $f$ is bijective, then $\#A = \#B$. This holds even when $A$ and $B$ have infinitely-many elements! These ideas can be used to make sense of when one "type" of infinity is bigger than another type of infinity.

The notion of different sizes of infinity was very controversial when first proposed by Georg Cantor in the late 19th century, but today is a commonly accepted and understood part of mathematics. For a very easy and brief introduction to the idea of different sizes of infinity, watch the short short TED-Ed video *How big is infinity?*,

https://youtu.be/UPA3bwVVzGI.

When a function $f : A \to B$ is bijective, there is always a function $g : B \to A$ which "undoes" $f$ in the following sense: for every $a \in A$, $g(f(a)) = a$, and for every $b \in B$, $f(g(b)) = b$. We call the function $g$ the ***inverse*** of $f$ and usually denote it by $f^{-1}$. (Notice that $f^{-1}$ *is not* $f$ raised to the negative first power! This is simply a common, if unfortunate, notation for the inverse.)

**Example 3.12.**
The bijective function $f$ is denoted in black in the image below, while its inverse $f^{-1}$ is given in red.




## 3.4 Images and preimages

Just as a function $f : A \to B$ associates elements of $B$ to elements of $A$, it also associates subsets of $B$ to subsets of $A$ by applying $f$ to every element

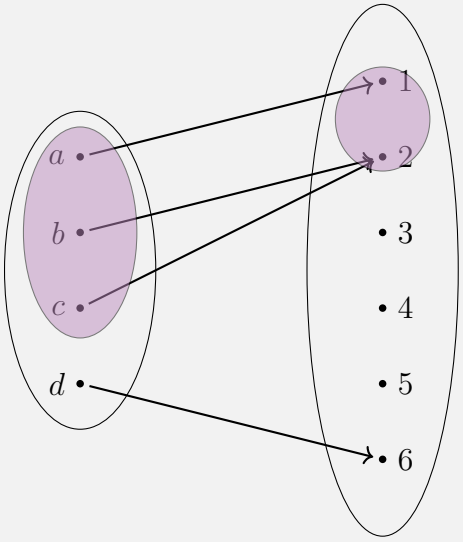
of a subset of $A$.

Suppose that $f : A \to B$ is any function and that $X \subseteq A$ is any subset of $A$. We can define a subset of $B$, which we'll denote $f(X)$, in the following way:

$$f(X) = \{f(x) \mid x \in X\}$$

This set $f(X)$ is called the ***image*** of $X$ under $f$.

**Example 3.13.**
Let $A = \{a, b, c, d\}$, $B = \{1, 2, ..., 6\}$ and let $f$ be the function from Example 3.1. If $X = \{a, b, c\}$, then its image $f(X)$ is $\{1, 2\}$.




Given any $Y \subseteq B$, the ***preimage*** of $Y$ is the set of all elements in $A$ which get mapped to an element of $Y$. The preimage is often denoted $f^{-1}(Y)$, even if $f$ is not bijective.

$$f^{-1}(Y) = \{x \in A \mid f(x) \in Y\}$$

**Example 3.14.**

Let $A$, $B$, and $f$ be as in Example 3.13. If $Y = \{1, 2\}$, then $f^{-1}(Y) = \{a, b, c\}$.

**Exercise 3.1.**
Let $A$, $B$, and $f$ be as in Example 3.13. What is the preimage of $\{3, 4, 5\}$?

## 3.5 Practice problems

**Problem 3.1.**
For each function $f$ given below, determine the domain and range of the function, and determine whether the function is injective, surjective, neither, or both. In these functions we are assuming the domain is a subset of the real numbers, and the codomain is $\mathbb{R}$.

(a) $f(x) = x^2$

(b) $f(x) = x^3$

(c) $f(x) = \sin(x)$

(d) $f(x) = \cot(x)$

(e) $f(x) = \sqrt{x}$

# Part II

# Basic Probability Theory

# 4 Basic Notions and Definitions

*It is likely that unlikely things should happen.*

ARISTOTLE

The outcome of many different types of events can not be predicted perfectly. For example, whether it will rain or snow tomorrow is never known for sure until it actually starts to precipitate. However, even if we can not know the outcome ahead of time with absolute certainty, we may still be able to measure the likelihood an event will occur. For instance, might be able to say there's a 75% chance it will rain tomorrow; or there is a 60% chance a given politician will win an elected office; or a 0.0000003% chance of winning the PowerBall Jackpot. In each situation we don't know for sure what will happen (maybe it will rain, maybe it won't; maybe the politician will be elected, but maybe not; maybe you'll buy the winning lottery ticket, but maybe you won't), however we may be able to assign a numerical value which measures how likely or unlikely each event is.

The mathematics used to assign these measures of likelihood belongs to the discipline of *probability theory*, and having a good understanding of the basics of probability theory will be required for everything else in this course. We will thus spend a fair amount of time slowly defining the basic definitions and ideas behind probability theory before we do anything else, starting "from the ground up" assuming you have never seen any probabilities before, and gradually introducing new, more involved ideas as necessary.

## 4.1 Experiments, sample spaces, and events

In the context of probability theory, an ***experiment*** is simply something whose outcome is not known with complete certainty. For example, flipping a coin, rolling a die, drawing a card from a shuffled deck, and recording the amount of time until a fish bites a hook are all experiments. In each case there are multiple possible outcomes (the coin may come up heads or tails; the die may roll a 1, 2, 3, 4, 5, or 6; the card may be the Ace of Spades, the King of Diamonds, the Two of Clubs, etc; the fish may bite the hook after one minute, or 39 minutes, or two hours, and so on), but we don't know exactly which outcome will take place ahead of time.

The collection of all possible outcomes of a given experiment is called the *sample space* of that experiment, and we will often denote the sample space of an experiment by the capital Greek letter omega, $\Omega$.

**Example 4.1.**
In the experiment of flipping a coin where the coin may come up heads or tails, we may use the symbols $H$ to mean the coin comes up heads, and $T$ to mean the coin comes up tails. The sample space of this experiment is then $\Omega = \{H, T\}$.

**Example 4.2.**
In the experiment where we draw a card from a shuffled deck of fifty-two standard playing cards, the sample space is the set of all cards,

$$\Omega = \{\text{Ace of Spades}, 2 \text{ of Spades}, 3 \text{ of Spades}, ...\}.$$

**Example 4.3.**
On the game show *Wheel of Fortune*, contestants try to solve word puzzles one letter at a time until they can determine the phrase that solves the puzzle. On their turn, each play must spin a wheel which determines the amount of money they win for correctly guessing a letter in the phrase, or if instead of winning money something "bad" happens (such as losing a turn or going bankrupt and losing all money that has been won).

If our experiment is the result of one spin of the wheel, then the sample space is the set of all possible results of one spin of the wheel, which is

$$\Omega = \{\text{Bankrupt}, \text{Lose a Turn}, \$500, \$600, \$650, \$700, \\ \$800, \$900, \$2500, \$1,000,000\}.$$

An *event* is a subset of the sample space of an experiment, and could consist of a single possible outcome (this is sometimes called a ***simple event***) or multiple possible outcomes (this is a ***compound event***).

**Example 4.4.**
In the example of rolling a six-sided die, so the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, depending on the number of dots (called *pips*) that appear when the die is rolled, the event where we roll a particular number like 5 is a simple event, $\{5\}$. The event where we roll one of several possible numbers, such as rolling an even number, is a compound event: $\{2, 4, 6\}$.

**Example 4.5.**
If our experiment is drawing a single card from a deck of shuffled playing cards, drawing one particular card such as the Ace of Spades, is a simple event: {Ace of Spades}. Drawing one of several possible cards, for example drawing a heart, is a compound event:

$$\{2 \text{ of Hearts}, 3 \text{ of Hearts}, ..., \text{Queen of Hearts}, \text{King of Hearts}\}.$$

**Example 4.6.**
Suppose a certain game involves rolling two distinguishable six-sided dice simultaneously (say, for example, one die is blue and one is red). What is the sample space? What event corresponds to the sum of the dice equalling eight?

The sample space consists of all possible rolls of the two dice. We'll denote the possible rolls as ordered pairs of numbers, where the first number corresponds to what value is rolled on the blue die, and the second roll corresponds to the value on the red die. For instance, $(3, 5)$ means a 3 appeared on the blue die and a 5 appeared on the red die. The sample space then consists of all thirty-six possible rolls of the

two dice:

$$\begin{aligned}\Omega = \{&(1,1),\ (1,2),\ (1,3),\ (1,4),\ (1,5),\ (1,6),\\ &(2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6),\\ &(3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6),\\ &(4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6),\\ &(5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6),\\ &(6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6)\}\end{aligned}$$

The event where the two dice sum up to eight is

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

In what follows, we will always let the sample space $\Omega$ of an experiment be the universal set so that we can talk about complements. (We'll see why this will be useful soon.)

Let's now make a three simple observations about events $E$ in a sample space $\Omega$:

1. The sample space $\Omega$ is itself an event, since $\Omega \subseteq \Omega$.
2. For any event $E \subseteq \Omega$, its complement $E^c \subseteq \Omega$ is also an event.
3. Given any sequence of events, $E_1$, $E_2$, $E_3$, ... (i.e., each $E_i \subseteq \Omega$ is itself an event), the infinite union of these events is itself an event:

$$\begin{aligned}\bigcup_{i=1}^{\infty} E_i &= E_1 \cup E_2 \cup E_3 \cup \cdots \\ &= \left\{\omega \in \Omega \,\middle|\, \omega \in E_i \text{ for some } E_i\right\}\end{aligned}$$

These three observations may not seem very interesting or important right now, but they are necessary in order to make probability theory rigorous. (We won't try to make probability theory too rigorous in this class, but it's good to know there is a way to do so.)

**Remark.**
Long ago, a lot of mathematics was done in a very loose and handwavy sort of way, but in the late 19th century and early 20th century

there was a big push to put a lot of the hand-wavy mathematics people had been using for hundreds of years on firm ground using principles of formal logic. Even though people had considered probability mathematically since at least the 16th century, probability theory was not put on firm logical ground until the middle of the 20th century.

Going through the details of making probability theory rigorous requires a rather technical detour through a branch of mathematics called *measure theory*, and if you continue to study probability and statistics for long enough – in particular if you studied these topics in graduate school – you would eventually see the measure-theoretic foundations we are skipping over.

## 4.2 Probability

The goal of probability theory is to associate numbers to events in such a way that these numbers indicate the likelihood the event will take place. An experiment results in only one outcome, but an event may consist of multiple possible outcomes (this is exactly what it means to be a compound event). In many situations we may only care that one of several possible outcomes took place, and not care about which particular output occured. For example, in playing a certain type of card game you may know that you'll win the game if the next card you receive has a value of 9 or higher, regardless of the suit. The event that you care about is then

$$\{9\heartsuit, 9\clubsuit, 9\diamondsuit, 9\spadesuit, 10\heartsuit, 10\clubsuit, 10\diamondsuit, 10\spadesuit, J\heartsuit, J\clubsuit, J\diamondsuit, J\spadesuit,$$
$$Q\heartsuit, Q\clubsuit, Q\diamondsuit, Q\spadesuit, K\heartsuit, K\clubsuit, K\diamondsuit, K\spadesuit, A\heartsuit, A\clubsuit, A\diamondsuit, A\spadesuit\}$$

and not which particular card you received.

To each event $E$ in the sample space $\Omega$ we are going to associate a number called the ***probability*** of the event $E$, and denoted $\Pr(E)$. We can think of this as a function Pr whose inputs are events $E$ (i.e., subsets of $\Omega$) and whose outputs are real numbers. To be considered a probability, though, we require that three conditions are satisfied:

1. For every event $E$, $0 \leq \Pr(E) \leq 1$.
2. $\Pr(\Omega) = 1$.

3. If $E_1$, $E_2$, $E_3$, ... is a (countably) infinite sequence of pairwise disjoint events (i.e., if each $E_i \cap E_j = \emptyset$ as long as $i \neq j$), then

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i).$$

These three conditions are called the ***axioms of probability***, and any time we are discussing probabilities we will *always* assume these three conditions are satisfied.

At first glance the axioms of probability may seem technical, so let's spend a minute thinking about what they actually mean.

1. A *probability* is supposed to represent a likelihood, and we are adopting the convention that probabilities close to zero mean the event is unlikely, and probabilities close to one mean the event is more likely. Sometimes instead of probabilities people talk about percentages, and you can convert probabilities to percentages by multiplying by 100. E.g., a probability of 0.78 corresponds to the percentage 78%. We don't want our likelihoods to be less than 0% or greater than 100% – it doesn't mean anything to say there a $-13\%$ chance something happens, or a 137% chance something happens – and so this means our probabilities need to be no less than 0 and no greater than 1.

2. When you perform an experiment *something* must happen, whatever it is. Since the sample space $\Omega$ represents all possible outcomes, some element of $\Omega$ must be the result of our experiment. That is, it's a sure thing that the event $\Omega$ will take place – there is a 100% chance (aka, probability 1) the event $\Omega$ will occur.

3. The third condition is maybe a little bit harder to see, but the idea is simple: if the event we care about can be broken up into lots of simpler pieces, we ought to be able to compute the probability of the original "large" event in terms of the probabilities of the simpler pieces. What we're claiming is that the right way to do this is to imagine chopping the big event into lots of non-overlapping, smaller pieces (this is the pairwise disjointness condition above), and then just add the probabilities of these smaller pieces together.

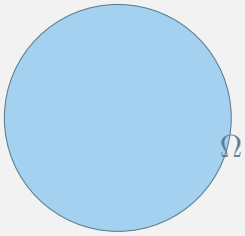
   The reason we want the events to not overlap (i.e., be pairwise disjoint) is because if they did overlap, then elements in the overlap would be counted multiple times. In principle this might be okay, we might be able to fix this by subtracting off some of the elements we

overcounted, but it makes the formula more complicated. (We'll see several concrete examples of this soon, so don't worry too much if that doesn't make complete sense right now.)

**Remark.**
There's actually a little white lie hiding in the axioms of probability above. Normally the way we'll want to calculate probabilities – i.e., the way we want to define the function Pr – won't satisfy all three axioms above for all possible subsets of the sample space $\Omega$. It's not at all obvious, but for lots of "natural" choices of $\Omega$ and probability function Pr you will get contradictory calculations (e.g., you could wind up saying things like $0 = 1$). Giving an example of this is a bit technical, but there's a nice explanation in the video *How the Axiom of Choice Gives Sizeless Sets* on YouTube (https://youtu.be/hcRZadc5KpI).

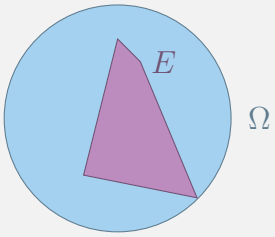
We can get around this issue by asking that the three axioms of probability not apply to *all* possible subsets of $\Omega$, but instead to a special collection of subsets called a $\sigma$-*algebra* (*sigma algebra*), which is just a fancy-sounding term for a collection of sets satisfying some nice properties.

**Example 4.7.**
Suppose you throw a dart at a round board – this is an experiment where the outcomes are the points on the board we can hit, but no matter how good you are at darts, you don't know with 100% certainty exactly which point on the board you'll hit. The sample space here is the collection of all points on the dart board,




The events of this experiment are subsets of the board, regions of the

board we may want to hit. (I.e., if you're playing an actual game of darts you may want to hit the "bullseye" in the center of the board to score 50 points, or land in a particular "wedge" of the board to score 15 points, etc.)




What is the probability a given $E$ takes place? That is, what is the probability a dart lands in a given region $E$? Assuming for simplicity that every point of the dart board is just as likely to be hit as every other point (e.g., maybe you're really bad at darts and it doesn't really matter where you aim at), then we claim that the correct probability function Pr is simply the area of the region $E$, divided by the area of the entire sample space (the dart board) $\Omega$:

$$\Pr(E) = \frac{\text{Area}(E)}{\text{Area}(\Omega)}.$$

Let's check that this really does satisfy the axioms of probability stated above:

1. Since areas are never negative, it's clear that for every $E$ we have $\Pr(E) = \frac{\text{Area}(E)}{\text{Area}(\Omega)} \geq 0$. As $E \subseteq \Omega$ (i.e., $E$ is contained inside of $\Omega$), the area of $E$ can't be any bigger than the area of $\Omega$: $\text{Area}(E) \leq \text{Area}(\Omega)$. Thus in our fraction for $\Pr(E)$, we must have that the numerator is no bigger than the denominator, and so $\Pr(E) \leq 1$.

2. Plugging $\Omega$ into our Pr function gives $\Pr(\Omega) = \frac{\text{Area}(\Omega)}{\text{Area}(\Omega)} = 1$.

3. The third property relies on a similar property for areas: if two regions in the plane, call them $E_i$ and $E_j$ for the moment, don't overlap, then $\text{Area}(E_i \cup E_j) = \text{Area}(E_i) + \text{Area}(E_j)$. If we have infinitely-many such non-overlapping regions, $E_1$, $E_2$, $E_3$, ..., this

extends to the infinite union:

$$\text{Area}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \text{Area}(E_i).$$

(If this feels a little hand-wavy, you can make it precise using properties of integrals, since the area of a region is equal to the double-integral of the constant function 1 over that region.)

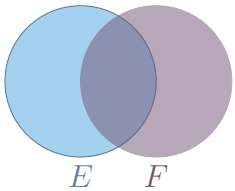
Plugging this into our formula for the Pr function, we can show the third axiom is satisfied:

$$\begin{aligned}
\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) &= \frac{\text{Area}\left(\bigcup_{i=1}^{\infty} E_i\right)}{\text{Area}(\Omega)} \\
&= \frac{\sum_{i=1}^{\infty} \text{Area}(E_i)}{\text{Area}(\Omega)} \\
&= \sum_{i=1}^{\infty} \frac{\text{Area}(E_i)}{\text{Area}(\Omega)} \\
&= \sum_{i=1}^{\infty} \Pr(E_i).
\end{aligned}$$

Now, keeping our dart board example above in mind, let's notice that if we had two events $E, F \subseteq \Omega$ which *did* overlap (i.e., they were not disjoint, $E \cap F \neq \emptyset$), then

$$\Pr(E \cup F) \neq \Pr(E) + \Pr(F)$$

because the region in the overlap, $E \cap F$, gets counted twice in the expression $\Pr(E) + \Pr(F)$: once for the $\Pr(E)$ term and once for the $\Pr(F)$ term. This is easy to see if we think of these regions as subsets of the plane and the probabilities as essentially the areas of the regions:




We'll see later that even if two events overlap we can express the probability of their union as a sum of the probabilities of the individual events,

but we'll have to compensate for this "double-counting" of the overlapping region. Before we explain how that formula works, it will be convenient if we discuss some simpler consequences of the probability axioms.

## 4.3 Consequences of the axioms

A common theme in advanced mathematics is to start off with a few simple assumptions, and then from those simple assumptions try to derive more useful conclusions. In Euclidean geometry, for example, there are five assumptions (usually called *the postulates*): every pair of points can be connected by a line segment; all line segments can be extended indefinitely in both directions; all right angles are congruent; given any two distinct points, there exists a circle whose center is one of the points and where the other point lives on the edge of the circle; given any line and a point not on that line, there is one line which goes through the given point and never intersects the given line. From these given simple assumptions, thousands of other geometric statements (such as the congruence theorems for triangles (side-angle-side, side-side-side, etc.), the Pythagorean theorem, the law of cosines, ...) can be deduced.

Similarly, from our three simple axioms of probability stated above, we can deduce many other useful statements. Here we go ahead and collect a few, starting with simple things and working our way up to more interesting statements.

**Remark.**
For the sake of completeness we will try to prove as many of these statements as possible. If you want to understand why a statement is true, then you should make an effort to sit down and understand the proof, and this is probably something you should really try to do if you're a math major.

However, you'll never be asked to regurgitate any of these proofs on a quiz or exam, so you *don't* need to try to memorize them. You definitely do need to know the statements of the propositions and theorems below, but you don't need to invest time and energy in memorizing the proofs.

Our first proposition is that the empty set must always have probability zero.

**Proposition 4.1.**
*For any experiment with sample space* $\Omega$, $\Pr(\emptyset) = 0$.

*Proof.*
Notice that $\emptyset = \emptyset \cup \emptyset \cup \emptyset \cup \cdots$, or written another way,

$$\emptyset = \bigcup_{i=1}^{\infty} \emptyset.$$

Furthermore all of these copies of the empty set are disjoint from one another: $\emptyset \cap \emptyset = \emptyset$. Thus our third axiom of probability tells us

$$\Pr\left(\bigcup_{i=1}^{\infty} \emptyset\right) = \sum_{i=1}^{\infty} \Pr(\emptyset).$$

Since $\emptyset = \bigcup_{i=1}^{\infty} \emptyset$, however, we can write this as

$$\Pr(\emptyset) = \sum_{i=1}^{\infty} \Pr(\emptyset).$$

On the right-hand side, notice that if $\Pr(\emptyset) > 0$, then $\sum_{i=1}^{\infty} \Pr(\emptyset)$ would be infinite since we would be adding some positive number to itself infinitely-many times. This would mean that $\Pr(\emptyset)$ is infinite, but this is impossible since probabilities are never bigger than 1. Hence the only possibility – the only option for $\Pr(\emptyset)$ that doesn't contradict the axioms – is that $\Pr(\emptyset) = 0$. □

At first glance the proposition above may not seem very interesting, but we're only using it as a stepping stone to more interesting things.

Our next consequence of the axioms says that in the third axiom, that the probability of an infinite disjoint union of events is the infinite sum of the probability of the individual events, applies also for finitely-many disjoint events. Before stating the proposition and giving the proof, let's realize that from the axioms we only know this statement for *infinitely-many* disjoint events. To get a comparable statement for finitely-many events, we

somehow need to turn a finite disjoint union into an infinite disjoint union and then apply the axiom. The trick for doing this will be to extend our finite list of events with infinitely-many copies of the empty set.

**Proposition 4.2.**
*If $E_1$, $E_2$, ..., $E_n$ is a finite collection of mutually disjoint events (i.e., if $E_i \cap E_j = \emptyset$ when $i \neq j$), then*

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_n) = \Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_n).$$

*Proof.*
We extend our initial finite list to an infinite list $E_1$, $E_2$, $E_3$, ... by setting $E_{n+1} = \emptyset$, $E_{n+2} = \emptyset$, $E_{n+3} = \emptyset$, and so on. Notice this is still a mutually disjoint collection of events, but now that we have infinitely-many events we can apply the third axiom. However, notice that since the union of any set $E$ and the empty set is just $E$, we have

$$\begin{aligned}\bigcup_{i=1}^{\infty} E_i &= E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n \cup \emptyset \cup \emptyset \cup \cdots \\ &= E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n.\end{aligned}$$

Thus when we apply the third axiom we have

$$\Pr(E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n) = \Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i).$$

If we write out the right-hand side, though, we have

$$\begin{aligned}&\sum_{i=1}^{\infty} \Pr(E_i) \\ &= \Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \cdots + \Pr(E_n) + \Pr(\emptyset) + \Pr(\emptyset) + \Pr(\emptyset) + \cdots\end{aligned}$$

By our previous proposition, we know $\Pr(\emptyset) = 0$, and so the sum above

becomes

$$\sum_{i=1}^{\infty} \Pr(E_i) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \cdots + \Pr(E_n).$$

Plugging this into the equation above we have the desired result:

$$\Pr(E_1 \cup E_2 \cup E_3 \cup \cdots \cup E_n) = \Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \cdots + \Pr(E_n)$$

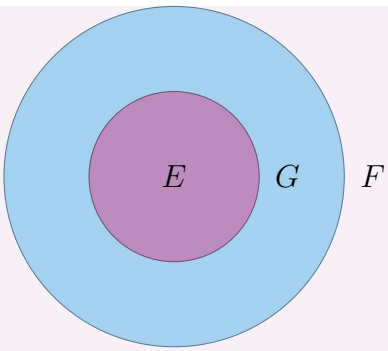
□

The above proposition will be used many, many times later in the course: it is our basic tool for breaking up complicated calculations into simpler ones.

Now we mention one more proposition which, at first glance, may seem like it should be obvious, but actually requires our previous propositions to prove.

**Proposition 4.3.**
*If* $E \subseteq F$, *then* $\Pr(E) \leq \Pr(F)$.

*Proof.*
Let $G$ be the set difference between $E$ and $F$ – i.e., $G$ consists of of everything "between" $E$ and $F$, $G = F \setminus E$. Then $F = E \cup G$ and $E \cap G = \emptyset$. Consider the picture below where *everything* in the big circle is an element of $F$, everything in the small circle is an element of $E$, and $G$ consists of the the ring between the big and small circles. Putting the small circle ($E$) and the ring ($G$) together, we reconstruct the original big circle ($F$), so $F = E \cup G$. Since there's no overlap between the small circle and ring, though, they are disjoint, $E \cap G = \emptyset$.




Since $E$ and $G$ are disjoint sets, $\Pr(E \cup G) = \Pr(E) + \Pr(G)$. Since $F = E \cup G$, though, we can write this as $\Pr(F) = \Pr(E) + \Pr(G)$. As $\Pr(G) \geq 0$ (probabilities are always non-negative), the probability fo $F$ is the probability of $E$ plus "a little bit more," and so $\Pr(F) \geq Pr(E)$. □

The idea that if $E$ is "smaller" than $F$ (i.e., $E \subseteq F$), then the probability of $E$ should be less than the probability of $F$ should seem reasonably intuitive. However, there is an odd consequence that the next example highlights.

**Example 4.8.**
Suppose we randomly select a point from the unit disc in the plane (or, equivalently, randomly throw a dart at a perfectly circular dart board that's one unit in radius). What is the probability we select the origin (i.e., the dart lands in the exact middle of the board)?

Before we compute this probability, let's notice that it is possible this could happen: it is conceivable we could by just random, dumb luck land right in the exact middle of the disc. The question we're trying to answer, though, is how likely is this to happen.

Notice that if $E$ is any little "subdisc" containing the origin, then the probability of landing at the origin must be less than the probability of $E$: the origin is a single point inside of $E$, so the set that contains just the origin is a subset of $E$, so by Proposition 4.3 the probability we hit the origin must be less than the probability we land in $E$.

However, the probability of landing in $E$ is just the area of $E$. By making this little disc $E$ smaller and smaller, we can make this proba-

bility arbitrarily small. For each of these little discs with small areas, the probability of selecting the origin must be even smaller. Since the probability of $E$ can be made arbitrarily small, the probability of landing at the origin must be smaller than every arbitrarily small number. The only possibility, then, is that the probability of selecting that one single point is zero.

**Remark.**
Notice that the probability in the previous example is zero, even though the event in question can still happen. That is, having probability zero *does not* mean an event is impossible! Rather, the probability is so astonishingly small – the event is so profoundly unlikely – that we can't assign any positive number to it.

We have one more important consequence of the axioms to consider before we take a break from stating propositions and look at concrete examples and compute probabilities. The next proposition is much more useful than you might think at first: as we will see, it will sometimes give us a trick for turning difficult calculations into easier ones.

**Proposition 4.4.**
*For any event $E$, the probability of the complement $E^c$ is one minus the probability of $E$:* $\Pr(E^c) = 1 - \Pr(E)$.

*Proof.*
Notice for any event $E \subseteq \Omega$, we have $\Omega = E \cup E^c$ and $E \cap E^c = \emptyset$. Thus $\Pr(\Omega) = \Pr(E \cup E^c) = \Pr(E) + \Pr(E^c)$. One of the axioms of probability stated, however, that $\Pr(\Omega) = 1$. Plugging this into the

equation above and solving for $\Pr(E^c)$ gives us our desired result:

$$\begin{aligned} \Pr(\Omega) &= \Pr(E) + \Pr(E^c) \\ \implies 1 &= \Pr(E) + \Pr(E^c) \\ \implies 1 - \Pr(E) &= \Pr(E^c) \end{aligned}$$

□

The above proposition is surprisingly useful: it says that computing the probability of an event's complement is basically just as good as computing the probability of the original event. That may not sound very helpful for interesting, but there are some problems where it's easier to think in terms of the complement.

**Example 4.9.**
Suppose we flip a coin ten times. What is the probability that a head appears at least once during those ten flips?

At first glance this looks like a pretty hard problem based on what we've discussed thus far, but it becomes very easy if we think about the complement. Say $H$ is the event we flip at least one heads (whether it's one head, two heads, three heads, ...). We want to find $\Pr(H)$, which seems hard to do directly, but maybe finding the probability $\Pr(H^c)$ is easier.

If $H$ is the probability we flip at least one heads, then $H^c$ is the probability we flip no heads. That is, we would have to flip tails on each of our ten flips. There are 1024 ways we can flip a coin ten times (i.e., 1024 different sequences of heads and tails – don't worry if you don't know where this number came from, we'll explain how to count things like this soon, but for right now I'm just telling this to you as a fact), but only one of those corresponds to flipping all tails. That is,

$$\Pr(H^c) = \frac{1}{1024}.$$

By our proposition above, this means

$$\Pr(H) = 1 - \Pr(H^c) = 1 - \frac{1}{1024} = \frac{1023}{1024} \approx 0.99902.$$

So there's about a 99.902% chance we would flip at least one head if we flipped a coin ten times.

**Exercise 4.1.**
Show that for any event $E$, $\Pr(E) = 1 - \Pr(E^c)$.

In this section we've mostly stated and proved some propositions, but haven't used them to do any calculations. In the next section we'll switch gears and focus on doing several examples, and in the process we will need to use the propositions above.

## 4.4 Examples

In the simplest possible examples, the sample space $\Omega$ consists of finitely-many events and each event is equally likely. The axioms of probability will be satisfied then only if the probability of each simple event (any one particular outcome) is one over the number of elements in the sample space.

To be more precise, suppose $\Omega$ contains $n$ elements:

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

Let's suppose that each element is equally likely. That is, for some value of $c$ which we will determine in a minute, we have $\Pr(\{\omega_i\}) = c$ for each $\omega_i \in \Omega$. Writing $\Omega$ as the disjoint union of sets which each contain one element,

$$\Omega = \{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_n\},$$

we must have

$$\Pr(\Omega) = \Pr(\{\omega_1\}) + \Pr(\{\omega_2\}) + \cdots + \Pr(\{\omega_n\}).$$

We know that $\Pr(\Omega) = 1$, however, and that each $\Pr(\{\omega_i\}) = c$. The above then becomes

$$1 = nc.$$

Solving for $c = 1/n$, we have that the probability of each event must by $1/n$.

**Remark.**
The discussion above is admittedly a verbose and pedantic way of stating something that should seem obvious. The point of the discussion, though, is simply to illustrate that we can always boil everything down to the axioms; the axioms of probability justify what we might intuitively guess.

**Example 4.10.**
Suppose a standard deck of fifty-two poker cards is shuffled and you draw a random card. What is the probability you draw $3\diamondsuit$, the three of diamonds?

Supposing that all cards are equally likely, since there are 52 cards in the deck and we know that the probability of drawing *some* card is 1 (by the axiom of probability which stated $\Pr(\Omega) = 1$), each card must have probability $1/52$. In particular, the probability of drawing the three of diamonds is $1/52$.

We can combine the observation above, that the probability of a simple event when the sample space contains $n$ elements and all are equally likely, with the propositions and axioms from earlier to compute probabilities of compound events.

**Example 4.11.**
Again suppose one card is drawn from a shuffled deck of fifty-two poker cards. What is the probability of drawing a diamond, regardless of its rank? (I.e., the probability we draw $2\diamondsuit$ or $3\diamondsuit$ or $4\diamondsuit$, ...)

Let's let $E$ denote the event that we draw a diamond. Notice there are thirteen different diamonds in the deck, and we can write this event as

$$E = \{2\diamondsuit, 3\diamondsuit, 4\diamondsuit, ..., 10\diamondsuit, J\diamondsuit, Q\diamondsuit, K\diamondsuit, A\diamondsuit\}.$$

We can then break up this one "complicated" event into several smaller ones; let $E_i$ be the simple event containing just the $i$ of diamonds,

whatever $i$ is. That is, we write

$$\begin{aligned} E &= E_2 \cup E_3 \cup E_4 \cup \cdots \cup E_{10} \cup E_J \cup E_Q \cup E_K \cup E_A \\ &= \{2\diamondsuit\} \cup \{3\diamondsuit\} \cup \{4\diamondsuit\} \cup \cdots \cup \{10\diamondsuit\} \cup \{J\diamondsuit\} \cup \{Q\diamondsuit\} \cup \{K\diamondsuit\} \cup \{A\diamondsuit\} \end{aligned}$$

Now we can use Proposition 4.2 to compute the probability of $E$:

$$\begin{aligned} \Pr(E) &= \Pr(E_2 \cup E_3 \cup \cdots \cup E_Q \cup E_K \cup E_A) \\ &= \Pr(E_2) + \Pr(E_3) + \cdots + \Pr(E_Q) + \Pr(E_K) + \Pr(E_A) \\ &= 1/52 + 1/52 + \cdots + 1/52 + 1/52 + 1/52 \\ &= 13/52 \\ &= 1/4 \end{aligned}$$

In general, assuming the sample space is finite and each simple event is equally likely, the probability of *any* event is the cardinality of the event divided by the cardinality of the entire sample space.

**Proposition 4.5.**
*If $\Omega$ is the sample space of an experiment with finitely-many possible outcomes and every outcome (simple event) is equally likely, then for any event $E$*

$$\Pr(E) = \frac{\#E}{\#\Omega}.$$

*Proof.*

Simply note that if $E = \{e_1, e_2, ..., e_n\}$, then

$$\begin{aligned}
\Pr(E) &= \Pr\left(\bigcup_{i=1}^{n}\{e_i\}\right) \\
&= \sum_{i=1}^{n} \Pr(\{e_i\}) \\
&= \sum_{i=1}^{n} {}^{1}/_{\#\Omega} \\
&= {}^{n}/_{\#\Omega} \\
&= {}^{\#E}/_{\#\Omega}
\end{aligned}$$

□

**Example 4.12.**
Suppose an urn contains sixteen marbles, eight of which are blue, six of which are green, and two of which are red. If you reach into the urn and randomly select one marble, what's the probability you select a green marble?

Here it's helpful if we imagine that the blue marbles are distinguishable – i.e., there's some way of telling the blue marbles apart and so we can so one of them is the first blue marble, one's the second blue marble, and so on. We do the same for the green and blue marbles.

We can then imagine the sample space as being

$$\Omega = \{b_1, b_2, ..., b_8, g_1, g_2, ..., g_6, r_1, r_2\}$$

where $b_1$ is the first blue marble, $b_2$ is the second blue marble, and so on.

Let's let $G$ denote the event where we get some green marble, whichever one it is. Then

$$G = \{g_1, g_2, g_3, g_4, g_5, g_6\}.$$

The probability we select a green marble, whichever one we happen to get, is then

$$\Pr(G) = \frac{\#G}{\#\Omega} = \frac{6}{16} = \frac{3}{8}.$$

The argument for the example above may seem like overkill: it seems completely obvious that if there are 16 marbles in the urn and 6 are green, then the probability of selecting a green one is $6/16$. However, not all the problems we are going to do are going to have "obvious," intuitive answers. Worse yet, sometimes the "obvious" answer is actually incorrect. Hence it's important for us to have a way of making our reasoning about calculating probabilities very precise – i.e., we need to know how to boil problems down to where we can apply the axioms or propositions we've developed. Right now we're showing how to do this with relatively straight forward examples just to get the basic ideas down, but will start doing more interesting examples soon.

**Example 4.13.**
Continuing with the same setup as in Example 4.12, what is the probability of drawing a non-green (either red or blue) marble?

We could do this by counting up all ways we could get a non-green marble (i.e., the number of ways to get a blue marble, plus the number of ways to get a red marble) and do something similar to what we did in finding the probability in Example 4.12, but we can also use Propsition 14 and note that getting a non-green marble is the complement of getting a green marble, so

$$\Pr(G^c) = 1 - \Pr(G) = 1 - 3/8 = 5/8.$$

In the next example we'll start to see why the issue of "overlapping" sets we alluded to before makes our calculations more complicated.

**Example 4.14.**
Suppose a survey is sent to 1000 students living in a residence hall at a university, and this survey contains two questions which are answered

simply as "yes" or "no." The questions are

1. Do you own a Playstation 4? Yes or no.
2. Do you own an Xbox One? Yes or no.

Suppose that of the 1000 students, 350 indicated they own a Playstation 4 (regardless of whether or not they also owned an Xbox One), 475 indicated they own an Xbox One (regardless of whether or not they also owned a Playstation 4), and 100 owned both.

Given the information above, if you pick a student at random, what is the probability that you will pick a student that owns at least one console?

Let's use $P$ to denote the set of students owning a Playstation 4 and $X$ the set of students owning an Xbox. Then $P \cup X$ is the set of students owning at least one console, and so we want to compute $\Pr(P \cup X)$. If these sets were disjoint we could simply compute $\Pr(P) + \Pr(X)$ and we'd be done. However, the sets are not disjoint, and so students owning both consoles would get counted twice (once for $\Pr(P)$ and once for $\Pr(X)$). That is, the students in $P \cap X$ are the ones that get counted twice.

One way to fix this would be to try to set everything up so that we *did* have a disjoint union. If we let $B$ be the set of students that owned both consoles, so $B = P \cup X$, then we could write

$$P \cup X = \underbrace{(P \setminus B)}_{\text{Own only PS4}} \cup \underbrace{(X \setminus B)}_{\text{Own only Xbox One}} \cup \underbrace{(B)}_{\text{Own Both}}.$$

We could then compute the probability with our rule that says we can add probabilities of disjoint unions:

$$\begin{aligned} \Pr(P \cup X) &= \Pr(P \setminus B) + \Pr(X \setminus B) + \Pr(B) \\ &= {}^{250}/_{1000} + {}^{375}/_{1000} + {}^{100}/_{1000} \\ &= {}^{725}/_{1000}. \end{aligned}$$

The example above showed one way to calculate probabilities where there's an overlap, but there is another way using a formula called the *inclusion-exclusion formula*. Intuitively, the inclusion-exclusion formula will say "if we overcounted by adding something twice, then subtract off one over overcounted terms." To make everything as precise as possible, we'll prove this intuitive-sounding proposition with the help of two simpler lemmas.

**Remark.**
A *lemma* is like a stepping stone for proving a more interesting proposition or theorem.

**Lemma 4.6.**
*For any two sets $E$ and $F$,*

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

*and these two sets, $E$ and $F \setminus E$, are disjoint.*

*Proof.*
We simply write out the definition of $E \cup F$ and rewrite the pieces bit-by-bit:

$$\begin{aligned} E \cup F &= \{x \,|\, x \in E \text{ or } x \in F\} \\ &= \{x \,|\, x \in E, \text{ or } x \in F \text{ but } x \notin E\} \\ &= \{x \,|\, x \in E \text{ or } x \in F \setminus E\} \\ &= E \cup (F \setminus E). \end{aligned}$$

Now we just need to verify these are disjoint:

$$\begin{aligned} E \cap (F \setminus E) &= \{x \,|\, x \in E \text{ and } x \in F \setminus E\} \\ &= \{x \,|\, x \in E, \text{ and } x \in F \text{ but } x \notin E\} \\ &= \emptyset \end{aligned}$$

□

The above lemma was completely about sets and had nothing to do with probability, but we can use it to prove a convenient fact about probabilities.

**Lemma 4.7.**
*If $E$ and $F$ are events in some sample space $\Omega$ and if $E \subseteq F$, then*

$$\Pr(F \setminus E) = \Pr(F) - \Pr(E).$$

*Proof.*
Notice that since $E \subseteq F$, $F = F \cup E$. (Intuitively, $F$ already contains $E$, so "adding" $E$ to $F$ with the union doesn't actually add anything.) Now we can use the lemma above to turn this into a disjoint union: $F \cup E = E \cup (F \setminus E)$. Keeping in mind $F = F \cup E$, however, we have

$$\begin{aligned} &\Pr(F) = \Pr(E) + \Pr(F \setminus E) \\ \implies &\Pr(F) - \Pr(E) = \Pr(F \setminus E). \end{aligned}$$

□

And now we are in a position to prove the inclusion-exclusion formula, which we will need to use at various points throughout the course.

**Proposition 4.8** (Inclusion-Exclusion)**.**
*For any two events $E$ and $F$ in a sample space $\Omega$, whether they are disjoint or not,*

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

*Proof.*
We simply use both lemmas above. First we write $E \cup F$ as a disjoint union $E \cup (F \setminus E)$; since these are disjoint the probabilities of each add together; we then rewrite $F \setminus E$ as $F \setminus (E \cap F)$; since $E \cap F \subseteq F$ we then use the second lemma above to turn this into a difference of

probabilities.

$$\begin{aligned}\Pr(E \cup F) &= \Pr(E \cup (F \setminus E)) \\ &= \Pr(E) + \Pr(F \setminus E) \\ &= \Pr(E) + \Pr(F \setminus (E \cap F)) \\ &= \Pr(E) + \Pr(F) - \Pr(E \cap F).\end{aligned}$$

□

**Example 4.15.**
Applying inclusion-exclusion to our example of students with Playstations and Xbox's (What's the correct plural of *Xbox*? Xboxes? Xboxs?) We have

$$\begin{aligned}\Pr(P \cup X) &= \Pr(P) + \Pr(X) - \Pr(P \cap X) \\ &= {}^{350}/_{1000} - {}^{475}/_{1000} - {}^{100}/_{1000} \\ &= {}^{725}/_{1000}\end{aligned}$$

**Example 4.16.**
Suppose a card is drawn from a shuffled deck of 52 playing cards. What is the probability the drawn card is a King or a Heart?

Notice there is overlap between the set of Kings and the set of Hearts (because there's a King of Hearts). But inclusion-exclusion applies even when there's overlap.

Letting $K$ denote the event we draw a King,

$$K = \{K\spadesuit, K\heartsuit, K\clubsuit, K\diamondsuit\}$$

and $H$ the event we draw a Heart,

$$H = \{2\heartsuit, 3\heartsuit, ..., 10\heartsuit, J\heartsuit, Q\heartsuit, K\heartsuit, A\heartsuit\}$$

we want to find the probability of getting a King *or* a Heart – this means the event we're interested in is the union $K \cup H$, whose probability we can easily calculate with inclusion-exclusion:

$$\begin{aligned}\Pr(K \cup H) &= \Pr(K) + \Pr(H) - \Pr(K \cap H)\\ &= {}^{4}\!/_{52} + {}^{13}\!/_{52} - {}^{1}\!/_{52}\\ &= {}^{16}\!/_{52}\\ &= {}^{4}\!/_{13}.\end{aligned}$$

(Here we used the following standard facts: there are four Kings in the deck; there are thirteen Hearts in the deck; and there is only one King of Hearts.)

**Example 4.17.**
Suppose a card is drawn from a shuffled deck of 52 playing cards. What is the probability the card is either a face card (Jack, Queen, or King), or a Club?

Let's let $C$ be the event we draw a club,

$$C = \{2\clubsuit, 3\clubsuit, ..., 10\clubsuit, J\clubsuit, Q\clubsuit, K\clubsuit, A\clubsuit\},$$

and $F$ the event we draw a face card,

$$\begin{aligned}F = \{&J\heartsuit, J\clubsuit, J\spadesuit, J\diamondsuit,\\ &Q\heartsuit, Q\clubsuit, Q\spadesuit, Q\diamondsuit,\\ &K\heartsuit, K\clubsuit, K\spadesuit, K\diamondsuit\}.\end{aligned}$$

Notice that $\#C = 13$, $\#F = 12$ and $\#(C \cap F) = 4$. By inclusion-exclusion, the probability of getting a Club or a face card (i.e., the even $C \cup F$) is

$$\begin{aligned}\Pr(C \cup F) &= \Pr(C) + \Pr(F) - \Pr(C \cap F)\\ &= {}^{13}\!/_{52} + {}^{12}\!/_{52} - {}^{4}\!/_{52}\\ &= {}^{21}\!/_{52}.\end{aligned}$$

**Example 4.18.**
Suppose a card is drawn from a shuffled deck of 52 playing cards. What is the probability the card is neither a face card (Jack, Queen, or King), nor a Club?

There are two ways we could do this problem: a hard way and an easy way. The hard way would be to count up all of the possible cards that are neither face cards nor Clubs. (Maybe this isn't actually hard, but it is tedious.) The better (easier) way to do this is to notice that getting neither a face card nor a Club is the complement of getting a face card or a club. That is, we want $\Pr((C \cup F)^c)$. Since we already know $\Pr(C \cup F)$ from the previous example, Proposition 14 tells us

$$\Pr((C \cup F)^c) = 1 - \Pr(C \cup F) = 1 - {}^{21}/_{52} = {}^{31}/_{52}.$$

**Example 4.19.**
Suppose an urn contains 16 marbles: 8 blue, 6 green, and 2 red. Suppose you randomly select two marbles one after the other without replacing the first marble. What is the probability *at least* one marble is blue?

Let's first just think about what the sample space $\Omega$ of this experiment is. If we were just drawing a single marble, the sample space would be something like

$$\{b_1, b_2, ..., b_8, g_1, g_2, ..., g_6, r_1, r_2\}.$$

How does this change if we draw two marbles?

We'll think of the two marbles as forming an ordered pair. E.g., we think of the pair $(b_3, r_2)$ as meaning we first pulled out the third blue marble, $b_3$, and then pulled out the second red marble, $r_2$.

If we replaced the first marble we drew before drawing the second one (so we could potentially draw the same marble twice), then the sample space would be the set of all pairs. Since we *are not* doing that, however, we have to remove those pairs where we have the same element twice.

We'll explicitly write the sample space out here and then soon see there's a better way to do this kind of calculation. So, the set of all

pairs of marbles where we don't have the same marble appear twice is

$$\{(b_1,b_2),(b_1,b_3),...,(b_1,b_8),(b_1,g_1),(b_1,g_2),...,(b_1,g_6),(b_1,r_1),(b_1,r_2),$$
$$(b_2,b_1),(b_2,b_3),...,(b_2,b_8),(b_2,g_1),(b_2,g_2),...,(b_2,g_6),(b_2,r_1),(b_2,r_2),$$
$$\vdots$$
$$(b_6,b_2),(b_6,b_3),...,(b_6,b_8),(b_6,g_1),(b_6,g_2),...,(b_6,g_6),(b_6,r_1),(b_6,r_2),$$
$$(g_1,b_1),(g_1,b_2),...,(g_1,b_8),(g_1,g_2),(g_1,g_3),...,(g_1,g_6),(g_1,r_1),(g_1,r_2),$$
$$(g_2,b_1),(g_2,b_2),...,(g_2,b_8),(g_2,g_1),(g_2,g_3),...,(g_2,g_6),(g_2,r_1),(g_2,r_2),$$
$$\vdots$$
$$(g_6,b_1),(g_6,b_2),...,(g_6,b_8),(g_6,g_1),(g_6,g_2),...,(g_6,g_5),(g_6,r_1),(g_6,r_2),$$
$$(r_1,b_1),(r_1,b_2),...,(r_1,b_8),(r_1,g_1),(r_1,g_2),...,(r_1,g_6),(r_1,r_2),$$
$$(r_2,b_1),(r_2,b_2),...,(r_2,b_8),(r_2,g_1),(r_2,g_2),...,(r_2,g_6),(r_2,r_1)\}$$

Let's take a minute to decipher what's happening above. We've tried to write out the sample space as a kind of table where the rows correspond to the choice of first marble, and the columns correspond to the choice of second marble, being careful to eliminate those pairs that would have the same marble picked twice (again, because we don't replace the first marble before picking the second).

Writing the sample space out in this table gives us a convenient way to figure out how big the sample space is. Let's jut notice there are 16 rows and 15 columns, so the size of the sample space is $\#\Omega = 16 \cdot 15 = 240$. (In case the 16 rows and 15 columns isn't clear, notice that we could choose any marble first and there are 16 marbles to choose from. For our second marble there are only 15 marbles we can choose since, no matter what marble we picked first, we can't pick the first marble twice.)

Now, our goal is to find the probability of getting at least one blue marble. We could sit down and go through our table above and manually count out how many pairs have at least one $b_i$ in them – but this is extremely tedious. A better way to answer the problem is to apply Proposition . Instead of computing probability directly, let's instead compute the probability of the complement.

The reason we're doing this is because the complement of a complement is the original set. (Sort of like how the negative of a negative cancels out to give you back the original value: $-(-x) = x$.) If $B$ is the event we get at least one blue marble, then $B^c$ would be the event

we do not get at least one blue marble – i.e., $B^c$ is the event we get no blue marbles. Using the fact $(B^c)^c = B$ and Proposition , we have $\Pr(B) = 1 - \Pr(B^c)$, so computing $\Pr(B^c)$ is basically just as good as computing $\Pr(B)$.

But, what is $\Pr(B^c)$? Well, we just said $B^c$ is the event we get no blue marbles, so how many ways could this happen? If we don't get a blue marble first, then there are eight possibilities for the first marble (six green plus two red). Now when we pick the second marble we want to pick one of the seven remaining non-blue marbles. This means $\#B^c = 8 \cdot 7 = 56$.

Putting all of this together, the probability we draw at least one blue marble is

$$\Pr(B) = 1 - \Pr(B^c) = 1 - \frac{56}{240} = \frac{184}{240} \approx 0.7666.$$

## 4.5 Limits of Events

The last bit of material in this section is a little more technical than what we've seen thus far, but will be useful later when we discuss random variables. We won't actually need the next few results for a while, so you could safely skip over this section of material on a first reading if you feel that it's too technical and then come back to it later when needed.

Suppose we have a non-decreasing sequence of events in our sample space. That is, we have a sequences of subsets $E_1$, $E_2$, $E_3$, ... of the sample space $\Omega$ with one additional requirement: to be non-decreasing we require

$$E_1 \subseteq E_2 \subseteq E_3 \subseteq \cdots$$

Notice that each event is contained in the next event. When we have a sequence like this, we can make sense of the limit of the probabilities of the events,

$$\lim_{n\to\infty} \Pr(E_n).$$

Notice this is really just a sequence of real numbers, and because the events are non-decreasing, the probabilities are also non-decreasing. I.e.,

$$\Pr(E_1) \leq \Pr(E_2) \leq \Pr(E_3) \leq \cdots$$

Now, let's notice that this sequence *must* have a limit because it is bounded above by 1.

**Proposition 4.9.**
*If $E_1$, $E_2$, $E_3$, ... is a non-decreasing sequence of events in a sample space $\Omega$, then*

$$\lim_{n\to\infty} \Pr(E_n) = \Pr\left(\bigcup_{n=1}^{\infty} E_n\right).$$

*Proof.*
The trick to proving this proposition is to rewrite the union $\bigcup_{n=1}^{\infty} E_n$ as a disjoint union. If we do this, then we can rewrite the probability of the union as the sum of probabilities, and this might be something we can more easily manipulate.

To turn the union above into a disjoint union, we want to split the $E_n$ events up into pieces that start off with $E_1$, then whatever we need to add to $E_1$ to get $E_2$, then whatever we need to add to $E_2$ to get $E_3$, and so on.

To do this, define $F_1 = E_1$ and for each $n > 1$ define $F_n = E_n \backslash E_{n-1}$. By construction, the $F_n$ are disjoint events and

$$E_n = F_1 \cup F_2 \cup \cdots \cup F_n.$$

Since this is a disjoint union, however, we can write

$$\Pr(E_n) = \Pr(F_1) + \Pr(F_2) + \cdots + \Pr(F_n).$$

If we also write $F = \bigcup_{n=1}^{\infty} F_n$, then $E = F$. Notice, however

$$\begin{aligned}
\Pr(E) &= \Pr(F) \\
&= \Pr\left(\bigcup_{n=1}^{\infty} F_n\right) \\
&= \sum_{n=1}^{\infty} \Pr(F_n) \\
&= \lim_{n\to\infty} \sum_{m=1}^{n} \Pr(F_n) \\
&= \lim_{n\to\infty} \Pr(F_1 \cup \cdots \cup F_n) \\
&= \lim_{n\to\infty} \Pr(E_n)
\end{aligned}$$

□

We have a similar proposition for limits of non-increasing events. Here we say a sequence of events $E_1$, $E_2$, $E_3$, ... is non-increasing if

$$E_1 \supseteq E_2 \supseteq E_3 \supseteq \cdots .$$

This means

$$\Pr(E_1) \geq \Pr(E_2) \geq \Pr(E_3) \geq \cdots$$

and so the sequence of probabilities is a non-increasing sequence of real numbers bounded below by zero, and hence it must have a limit. The next proposition says that we can find this limit as the probability of the intersection of the events.

**Proposition 4.10.**
*If $E_1$, $E_2$, $E_3$, ... is a non-increasing sequence of events in a sample space $\Omega$, then*

$$\lim_{n\to\infty} \Pr(E_n) = \Pr\left(\bigcap_{n=1}^{\infty} E_n\right).$$

**Exercise 4.2.**
Prove Proposition 4.10 by applying de Morgan's laws to turn the non-increasing sequence of events into a non-decreasing sequence and applying Proposition 4.9.

## 4.6 Practice problems

**Problem 4.1.**
Suppose a large bag of loose change contains the following coins: nine pennies, sixteen nickels, twelve dimes, and thirteen quarters. If you reach into the bag without looking and randomly pull out one coin, what is the probability you pull out a quarter?

**Problem 4.2.**
Suppose, as in the previous problem, a large bag of loose change contains the following coins: nine pennies, sixteen nickels, twelve dimes, and thirteen quarters. If you reach into the bag without looking and randomly pull out two coins, what is the probability that at least one of those coins is a quarter? (Hint: You can think that you pull the coins out one at a time, but after you pull the first coin out of the bag you *don't* put it back in the bag.)

**Problem 4.3.**
Suppose that two fair, six-sided dice are rolled. What is the probability that the sum of the dice is nine?

**Problem 4.4.**
Suppose a coin jar contains the following twenty coins: eight pennies, six nickels, four dimes, and two quarters. If you reach into the coin jar and randomly pull out four coins, what is the probability you pulled out *at least* ten cents?

⋆ **Problem 4.5.**
Alice and Bob play the following game: Alice and Bob take turns flipping a fair, two-sided coin. I.e., Alice flips, then Bob flips, then Alice flips, then Bob flips, ... The game continues until a head is flipped, and whoever flips the head is the winner of the game. If Alice goes first, what is the probability she wins the game?

# 5 *Counting*

*We do not want to count; we want to think counting.*

Alain Badiou
*Number and Numbers*

In the last chapter we saw that many probability problems boil down to counting: Proposition 4.5 said that if we had a finite sample space and all simple events were equally likely, then the probability of any compound event is the number of elements in the event divided by the number of elements in the sample space. Thus to answer these types of problems we need to be able to count the number of elements in the sample space, and the number of elements in a given event.

In particular, we want to be able to count in a very precise and systematic way. That is, we want to develop some tools that will turn what could be a very tedious to do by hand (e.g., counting the number of possible 7-character license plates) to something very simple (multiplying some numbers based on what characters are allowed in the license plate).

## 5.1 Permutations

Sometimes we may care about the total number of ways to order all of the elements in a set. For example, suppose that we have the following six letters

`A`, `E`, `L`, `R`, `S`, `W`

etched onto wooden tiles. If we were to put the tiles into a bag and randomly pull them out one at a time, what is the probability we'd pull them out in the order spelling the word `RAWLES`? To answer this we need to know the number of elements in the sample space – i.e., the number of arrangements of the six letters.

In general, if a set contains $n$ distinct elements (in the above, $n = 6$), the number of different arrangements of those $n$ elements is given by the product

$$n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ldots \cdot 4 \cdot 3 \cdot 2 \cdot 1.$$

The reason for this is that when we choose the first element of the arrangement, we can choose any element of the set; there are $n$ elements in the

set, so $n$ possible choices. When we choose the second element of the arrangement, we must choose from the *remaining* elements of the set. We've already used up one element when we chose the first entry in the arrangement, so we have one fewer, $n-1$, possibilities for the second element. Similarly, there are $n-2$ options for the third element, and so on, down until there's only one option (whatever happens to remain) for the very last element.

This quantity comes up many times in mathematics and is tedious to write out by hand each time, but luckily there's some notation for it. The product defined above is denoted $n!$ and is called $n$ ***factorial***. For example,

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

One convenient property of factorials is that they can be written in terms of factorials of smaller numbers. In particular, if you already know what $n!$ is, then it's easy to compute $(n+1)!$. By definition, $(n+1)!$ is the product

$$(n+1)! = (n+1) \cdot n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1.$$

Notice the $n$ factors appearing to the right of $(n+1)$ on the right-hand side above exactly give us $n!$:

$$(n+1)! = (n+1) \cdot \underbrace{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1}_{n!}.$$

Thus we can write

$$(n+1)! = (n+1) \cdot n!$$

or equivalently,

$$n! = n \cdot (n-1)!.$$

Thus if we've already calculated $5! = 120$, then it's easy to compute $6!$:

$$6! = 6 \cdot 5! = 6 \cdot 120 = 720.$$

**Example 5.1.**
Suppose six wooden tiles have the letters `A`, `E`, `L`, `R`, `S`, `W` etched on them. If these tiles are placed in a bag and randomly pulled out one-by-one, laying the tiles from left-to-right on a table as they're pulled out, what is the probability the tiles spell out `RAWLES`?

There are six tiles, so $6! = 720$ possible words we could spell out. Of these, only one is `RAWLES`, so the probability we spell out `RAWLES`

is $1/720$.

**Remark.**
By convention 0! is defined to be 1. Using the interpretation that $n!$ tells us the number of arrangements of $n$ things, this convention makes sense: there is one way to arrange zero things, the "empty" arrangement.

More generally, if we are constructing a finite ordered sequence of $r$ objects, let's momentarily call this sequence

$$\sigma_1 \, \sigma_2 \cdots \sigma_r$$

where $\sigma_i$ denotes the entry in the $i$-th position, the number of possible sequences is the product of the number of options for $\sigma_1$ times the number of options for $\sigma_2$ times the number of options for $\sigma_3$, and so on, up through the number of options for $\sigma_r$.

**Example 5.2.**
Suppose instead of using all of the letters from `A`, `E`, `L`, `R`, `S`, `W` we only use three letters. That is, we put all six of our tiles with these letters in a bag, and randomly pull out three, placing them in front of us from left-to-right as they are pulled out. How many three-letter sequences can we spell out?

There are still six options for the first letter, five options for the second letter, and four options for the third letter. This means there are

$$6 \cdot 5 \cdot 4 = 120$$

possible three-letter sequences we can spell out.

**Example 5.3.**
With the same setup as in Example 5.2, what is the probability the word SAW is spelt?

From Example 5.2 we know there are 120 distinct three-letter sequences we can construct, and SAW there's only one of them that spells SAW, so the probability of spelling out SAW is $1/120$.

**Exercise 5.1.**
Again suppose we randomly draw three tiles one at a time, without replacement, from a bag containing six tiles with the letters A, E, L, R, S, W.

(a) What is the probability we construct a three letter sequence that beings with the letter L?

(b) What is the probability we construct a three letter sequence whose first two letters are LA in that order?

(c) What is the probability our three letter sequence ends in W?

The idea above that we should just multiply the number of options for each entry in our sequence together works even if we allow repetition.

**Example 5.4.**
Suppose a four-digit code on a bike lock is made up of the digits zero through nine and digits may be repeated (e.g., 1122 is a valid combination). How many possible combinations are there? If you forgot your combination and randomly guessed a combination, what is the probability you'd guess the right combination?

For each digit in our combination there are ten options (0 through 9 gives ten possibilities). Hence the number of combinations is $10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = 10000$. Of these ten-thousand combinations, only one is the right one that opens the lock. If you randomly guessed one

combination, the probability it'd be the correct combination is $1/10000$.

So, moral of the story: don't forget the combination to your bike lock unless you want to manually enter ten-thousand combinations until you find the right one.

**Example 5.5.**
Suppose license plates in a certain state follow the following pattern: three letters followed by a dash followed by four digits. How many possible license plates are there? (Assume letters and digits may be repeated.)

For each of the letters we have twenty-six options because there are twenty-six letters in the English alphabet, `A` through `Z`. For the digits, we again have ten options each time (`0` through `9`). Thus the number of combinations is

$$26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 26^3 \cdot 10^4 = 175,760,000.$$

(Notice this is more than four times the number of people in the most populous state: California has about 40-million residents. So even if each resident of California owned four cars, there would still be enough license plates using the pattern above for each car to have a unique license plate.)

## 5.2 Combinations

Notice that in the examples we've seen thus far the order of our sequences mattered. For example, the bike lock combination `1327` is very different from the combination `7312`, even though they both have the same digits. But there are some problems where the order does not matter. For instance, maybe you have seven friends and you want to choose four of them with which to play frisbee. Here, the order in which you select the friends doesn't matter.

For example, say your seven friends are Alice, Bob, Cassandra, Eric, Danielle, Fred, and George. It doesn't matter if you choose to Alice, Eric,

Danielle and George versus George, Eric, Alice, and Danielle – you're still playing frisbee with the same people.

So, continuing with the frisbee example, how many different ways can you choose four of these seven friends to play frisbee with? If the order did matter, then we from what we discussed in the previous section there would be $7 \cdot 6 \cdot 5 \cdot 4$ ways to choose your four friends. However, this number is going to be too big when the order doesn't matter, and the reason it's too big is that you over-count. For example, the two orderings mentioned above (Alice, Eric, Danielle, George and George, Eric, Alice, Danielle) would be counted as two separate things, which is what we don't want.

So, how should we fix this over counting? Let's try to reason our way through what to do by considering a smaller example: let's say you only wanted to pick two of the seven friends. If order mattered, there would be $7 \cdot 6$ options. This is too big because Alice & Bob, for example, gets counted as something distinct from Bob & Alice, and likewise for any collection of two people. I claim this means that $7 \cdot 6$ is exactly twice as big as what we want: for each choice (e.g., Alice & Bob) there's an "equivalent" choice (Bob & Alice) that $7 \cdot 6$ counts. That is, $7 \cdot 6$ "sees" two options where we really only have one. To fix this, let's cut the number in half to get

$$\frac{7 \cdot 6}{2} = 21.$$

So there are 21 different ways we can select two friends from our seven friends above, if the order in which we select them doesn't matter.

**Exercise 5.2.**
Explicitly write out all of the ways to select two friends from the seven above if order doesn't matter to convince yourself there are twenty-one options.

Now, let's try to extend this to three friends. If order mattered there would be $7 \cdot 6 \cdot 5$ possibilities. Since order doesn't matter, however, we're overcounting. When there were two friends we were overcounting by a factor if 2 because there were two "arrangements" of the two chosen friends. For three friends, by how much is $7 \cdot 6 \cdot 5$ overcounting? Well, if we picked three friends – say Alice, Bob, Cassandra – there are $3! = 6$ ways we could arrange them:

Alice, Bob, Cassandra
Alice, Cassandra, Bob
Bob, Alice, Cassandra
Bob, Cassandra, Alice
Cassandra, Alice, Bob
Cassandra, Bob, Alice

these six arrangements are counted as distinct possibilities in our $7 \cdot 6 \cdot 5$ calculation, and something similar happens for any grouping of three friends that we pick: there will be six ways to arrange them which are counted as distinct in the $7 \cdot 6 \cdot 5$ calculation. That is, $7 \cdot 6 \cdot 5$ overcounts the ways we can choose three of our seven friends, if order doesn't matter, by a factor of 6. Hence the correct number of ways we can choose three friends from our seven is

$$\frac{7 \cdot 6 \cdot 5}{6} = 35.$$

Now we should see how this extends to choosing four friends: the calculation $7 \cdot 6 \cdot 5 \cdot 4$ includes the ordering of the friends, there are $4! = 24$ orderings for any grouping of four friends, so we need to divide this calculation by 24 to get

$$\frac{7 \cdot 6 \cdot 5 \cdot 4}{24} = 35.$$

**Remark.**
Notice the number of ways to choose four friends from seven is the same as the number of ways to choose three from seven. This isn't a coincidence: choosing which four friends you want to play frisbee with is the same as choosing which three friends you don't want to play frisbee with.

Extending the argument above to choosing $r$ elements from a collection of $n$ distinct things, where order does not matter, we see that the number of possibilities is

$$\frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+2) \cdot (n-r+1)}{r!}.$$

Let's compare this to the cases we've already discussed:

- If we were to choose two of our seven friends, then in the expression above we would have $n = 7$ and $r = 2$. In the numerator we would start the product at 7 and continue down to $n-r+1 = 7-2+1 = 6$, giving us just two factors. The expression then becomes

$$\frac{7 \cdot 6}{2!} = 21$$

  as before.

- If we were to choose three of our seven friends, then we would have $n = 7$ and $r = 3$. The numerator would start at 7 and continue down to $n - r + 1 = 7 - 3 + 1 = 5$. This gives us

$$\frac{7 \cdot 6 \cdot 5}{3!} = 35$$

- Choosing four of seven friends, $n = 7$, $r = 4$, and $n - r + 1 = 4$, so the expression is

$$\frac{7 \cdot 6 \cdot 5 \cdot 4}{4!} = 35.$$

This is a little bit tedious to write out, especially if $n$ is a very large number, but luckily there's some notation for this expression. To explain the notation, let's first make an observation. If we only want part of $n!$, say we only want the first $r$ factors, then what we can do is take $n!$ and divide out the factors we don't want. For example, say we only want the first four factors of 10!; i.e., $10 \cdot 9 \cdot 8 \cdot 7$. We can write this in terms of factorials by dividing out the portion of 10! that we don't want: we don't want everything from 6 down to 1, aka 6!, so we can just divid it out:

$$\frac{10!}{6!} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 10 \cdot 9 \cdot 8 \cdot 7.$$

Notice that the 6 in 6! equals $10 - 4$ and we wanted the first four factors of 10!.

In general, the first $r$ factors of $n!$,

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+2) \cdot (n-r+1)$$

can be written as

$$\frac{n!}{r!}.$$

For instance, the $7 \cdot 6 \cdot 5$ that appeared when we selected three of our seven friends can be written as $7!/(7-3)!$. Keeping in mind we then divided

this quantity by 3! to account for all of the ways we could arrange the three friends we selected, we see our expression above may be written as

$$\frac{\left(7!/(7-3)!\right)}{3!} = \frac{7!}{(7-3)!\,3!}$$

These types of expressions come up a lot in probability and other areas of mathematics, and so they have a special notation. The expression

$$\frac{7!}{(7-3)!\,3!}$$

is often written as

$$\binom{7}{3}$$

and is called *"seven choose three"*.

More generally, we define $\binom{n}{r}$ as the expression $\frac{n!}{(n-r)!\,r!}$ and call this $n$ ***choose*** $r$:

$$\binom{n}{r} = \frac{n!}{(n-r)!\,r!}.$$

Extending the discussion we had above about choosing friends to play frisbee with proves the following proposition:

**Proposition 5.1.**
*If $n \geq r \geq 0$, then the number of ways to choose $r$ elements from a collection of $n$ distinct elements where order does not matter is equal to $\binom{n}{r}$.*

Before reading any more in this section, you should try to do the computations in the following exercise to be sure you understand how to compute $\binom{n}{r}$.

**Exercise 5.3.**
Compute the following:

(a) $\binom{5}{3}$

(b) $\binom{9}{2}$

(c) $\binom{9}{7}$

(d) $\binom{37}{37}$

(e) $\binom{37}{0}$.

(The last two aren't as bad as you might think: just write out the expression and cancel out what you can and something nice will happen.)

The notation $\binom{n}{r}$ is convenient, but if you actually want to do the computation it can require a little bit of work since you have to compute factorials. It would be convenient, then, to know at least a couple of simple properties of $\binom{n}{r}$ which will sometimes make work a little bit easier. We'll state three of the properties as a proposition whose proof is left as an exercise, but it will be a straight-forward exercise *(that you should try to complete!)*: just write out the definitions of the quantities described by the proposition and cancel out what you can.

**Proposition 5.2.**
*If $n \geq 0$ is an integer, then we have the following:*

1. $\binom{n}{0} = 1$
2. $\binom{n}{n} = 1$
3. *For any integer $k$ with $n \geq k \geq 0$,* $\binom{n}{k} = \binom{n}{n-k}$.

**Exercise 5.4.**
Prove Proposition 5.2.

## 5.3 Examples

Now that we have some tools for counting, let's use them to answer some question.

**Example 5.6.**
How many five-card poker hands are there?

Here we imagine we have a shuffled deck of fifty-two playing cards and we select five of the cards without replacement. Let's notice that the order we select the cards in doesn't matter: if you're playing poker, you don't care if the first card you get is the King of Hearts and the second is the Ace of Spades, versus the first one being the Ace of Spades and the second one being the King of Hearts. The important thing is the collection of cards you have, not the order in which you received them.

There are fifty-two cards we can choose and we're choosing five of them, so the number of possible choices is $\binom{52}{5}$. If we do this computation on a calculator or computer (it's a bit too tedious to do by hand – 52! is a huge number), we'd find

$$\binom{52}{5} = 2,598,960.$$

So a little more than 2.5-million possible five-card poker hands.

**Example 5.7.**
How many ways are there to get a four-of-a-kind in a five-card poker hand? (A four-of-a-kind means we have all four cards of one rank, and then any other fifth card. For example, (7♣, 7♠, 7♡, 7♢, 3♣) is such a hand.)

So, if we're going to "construct" a four-of-a-kind, what information do we have to determine? Well, we have to determine which rank we'll have all four cards of (e.g., will we have all four two's, or all four three's, or all four four's, ...), and then we have to pick one more card.

There are thirteen ranks we can choose for the four cards where we have all four cards of that rank, so $\binom{13}{1} = 13$ ways we can select that

portion of our hand. And since we can choose any of the remaining forty-eight cards (there were fifty-two, but we've just up four of them), we have $\binom{48}{1} = 48$ choices for the remaining card.

Now we think of this information as having two parts: the rank of the four cards, and then the rank and suit of the one remaining card. We now know there are 13 possibilities for the rank, and 48 possibilities for the one remaining card, so the total number of four-of-a-kind's we can have is

$$13 \cdot 48 = 624.$$

**Example 5.8.**
What's the probability a random five-card poker hand will be a four-of-a-kind?

From the previous two examples we know there are $2,598,960$ possible five-card poker hands, but of these only 624 are are a four-of-a-kind. Hence the probability a random poker hand will be a four-of-a-kind is

$$\frac{624}{2598960} = 0.0002401.$$

That is, there's only a 0.02% chance a random hand is a four-of-a-kind. It is precisely because such a hand is so unlikely why this is considered a good hand in poker: the best poker hands are the ones that are the least likely.

In the examples above it has been clear, either because it was explicitly stated or obvious from the context, whether order mattered in our count or not. Sometimes this may not be quite so obvious, however, as the next example illustrates.

**Example 5.9.**

(a) Suppose you roll two distinguishable, fair six-sided dice – suppose one die is blue and one die is red – how many possible outcomes are there?

(b) Suppose you roll two indistinguishable, fair six-sided dice – say both are white – how many possible outcomes are there?

(a) Let's imagine that after we roll the dice, we record the outcome as an ordered pair $(b, r)$ where $b$ is the value on the blue die and $r$ is the value on the red die. There are six possibilities for $b$, and six possibilities for $r$, so the number of possible rolls for these two dice is $6 \cdot 6 = 36$.

(b) When rolling the two white dice, let's record the outcome of again as an ordered pair. But this time we'll call the value $(M, m)$ where $M \geq m$ – so, first we write down the larger value, then we write down the smaller value. (We can't distinguish the dice, so we can't say one of the dice is "first" like we did when we had the blue and red dice.)

Now how many possible values are there for $M$ and $m$? This is a little trickier than our earlier problems because what values are available for $m$ depends on what $M$ is. Notice that $M$ could take on any possible value, so there are six possibilies for $M$. However, once we know what $M$ is, there are $M$ possibilities for $m$: since $M \geq m$, $m$ must be one of 1, 2, ..., $M$.

Putting all of this together, we have

$$1 + 2 + 3 + 4 + 5 + 6 = 21$$

possible ways to roll two indisguishable dice.

**Exercise 5.5.**
Suppose you're considering possible European vacations, and there are eight different cities you're interested in visiting: Athens, Brussels, Coppenhagen, Dublin, Eindhoven, Florence, Glasgow, and Helsinki. However, you only have enough time and money to visit three of these eight cities.

Let's suppose an *itinerary* for your trip consists of an ordered list of the three cities you will visit, given by the order in which you visit the cities. For example, one itinerary might be *Brussels, Dublin, Athens.*

To save ourself some writing, let's just list this by the first letter of each city, so this itinerary is `BDA`. This means we first visit Brussels, then we go to Dublin, then finally visit Athens before heading home.

(a) How many possible three-city itineraries are there?

(b) How many itineraries include Eindhoven as the first city?

(c) How many itineraries include Helsinki, regardless of order?

(d) If we decided we didn't care about the order in which we visited the cities, just the cities we visit, how many possible trips are there?

(e) How many unordered trips include Helsinki?

**Example 5.10.**
How many ways are there to get a two-pair in a five-card poker hand?

(A two-pair consists of a hand where we have two cards of the same rank (such as two 2's), two cards of a different rank (e.g., two 7's), and then one card of yet a third rank (say King of Clubs). For example, $(2\heartsuit, 2\clubsuit, 7\diamondsuit, 7\heartsuit, K\clubsuit)$, is a two-pair.)

To build a two-pair we first need to choose the two different ranks which will appear in our pairs. There are 13 ranks and we choose 2, so there are $\binom{13}{2}$ ways to do this. Now for each of those ranks we have to choose two of the four possible suits, and there are $\binom{4}{2}$ ways to do this. Notice we need to do this twice, once for each pair. Finally, we choose one more card that can be any card as long as it's of a different rank. There are 11 remaining ranks and for each one we have 4 possible suits, so there are $\binom{11}{1} \cdot \binom{4}{1}$ ways to choose that last card. Putting all of this together, there are

$$\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot \binom{11}{1} \cdot \binom{4}{1} = 123552$$

possible two-pair hands.

**Example 5.11.**
How many ways are there to arrange the letters A, E, H, O, P, R, R, R, T, T, Y to spell distinct 11-letter sequences of the letters? I.e., if we swap two identical letters, this should still count as the same spelling. What is the probability a randomly constructed 11-letter sequence spells HARRYPOTTER?

If all of our letters were distinct, there would 11! possible arrangements. Since the letters are not distinct, however, this count is too big. The issue is that we are over-counting because in the 11! we are supposing swapping two different R's or T's is a different spelling, which it is not. To compensate for this we need to divide out all of the ways we could swap R's or T's and still spell the same word.

Once we've chosen the order of our letters, we have 3! ways we could re-arrange the R's to still spell the same word, and 2! ways to re-arrange the T's. Thus the number of distinct 11-letter sequences is

$$\frac{11!}{3!2!} = 3,326,400.$$

As for computing the probability we spell HARRYPOTTER, there are two ways we could approach the problem. We could notice that of all the distinct 11-letter words, only one spells out HARRYPOTTER and so the probability is

$$\frac{1}{\left(\frac{11!}{3!2!}\right)} = \frac{1}{3326400}.$$

Alternatively, we could consider the 11! arrangements assuming we can distinguish the R's and T's, but then count up the number of arrangements spelling HARRYPOTTER. Here the position the only H, the only A, Y, P, O, and E are all fixed: to spell out HARRYPOTTER these characters *must* be in positions one, two, five, six, seven, and ten, respectively. However, we can rearrange the three R's as much as we like among the second, third, and eleventh positions: there are 3! ways to do this. We can also swap which T's are in the eigth and ninth positions: there are 2! ways to do this. Hence the probability is

$$\frac{3!2!}{11!} = \frac{12}{39,916,800} = \frac{1}{3362400}.$$

**Example 5.12.**
Suppose a special three-character code is required by each user of a certain website. This code must consist of exactly one upper case letter and two digits. How many possible codes are there?

To do this problem we first think about choosing the "format" of the code. Using `U` as a placeholder for an uppercase letter and `D` as a placeholder for a digit, we see there are three possible formats: `UDD`, `DUD`, `DDU`. A more sophisticated way of counting this (instead of just writing down an exhaustive list of all possible formats) is to notice that our string specifying the format has three letters, but two of them are repeated because we have two digits. Dividing out the number of ways to swap the two `D`'s in the format, there are $\frac{3!}{2!} = 3$ possible formats.

Once we've chosen a format, we then choose the letter and two digits. (Notice the digits may repeat, but don't have to. This does not affect our count for the number of formats above. The two `D`'s above are just placeholders for arbitrary digits.) In each case there are $26 \cdot 10^2$ codes in the desired format, and so the total number of codes is

$$\frac{3!}{2!} \cdot 26 \cdot 10^2 = 7800.$$

An alternative way to think about this, if you're worried we might be over- or under-counting because we may or may not repeat the digits is to divide each format into two cases: one where the digit repeats and one where the digits are distinct. Let's compute the number of codes this way as well and see this gives the same value.

Here we'll break each format up into two cases. When the digits are distinct we will write `D` for the first digit and $\texttt{D}_2$ for the second digit, which must be different from the digit we have in `D`. When digits repeat we'll write `D` for the first digit, and use `R` as a placeholder for the second digit to indicate this second digit must be the same as the first digit. There are now six formats we consider:

$$\texttt{UDD}_2,\ \texttt{UDR},\ \texttt{DUD}_2,\ \texttt{DUR},\ \texttt{DD}_2\texttt{U},\ \texttt{DRU}.$$

For the formats with two distinct digits there are $26 \cdot 10 \cdot 9$ possible codes. (Since we have different digits the second digit can't match up with the first digit, so we have one fewer choice). For the formats with repeate digits there are $26 \cdot 10 \cdot 1$ possible codes. Adding all of these

possible codes together for our formats above gives

$$\begin{aligned}
&26 \cdot 10 \cdot 9 + 26 \cdot 10 \cdot 1 + 26 \cdot 10 \cdot 9 + 26 \cdot 10 \cdot 1 + 26 \cdot 10 \cdot 9 + 26 \cdot 10 \cdot 1 \\
=&26 \cdot 10(9+1) + 26 \cdot 10(9+1) + 26 \cdot 10(9+1) \\
=&26 \cdot 10^2 + 26 \cdot 10^2 + 26 \cdot 10^2 \\
=&3 \cdot 26 \cdot 10^2 \\
=&7800.
\end{aligned}$$

**Example 5.13.**
Suppose passwords for a website must satisfy the following conditions:

- Passwords consist of uppercase letters, lowercase letters, one of five symbols (`@`, `#`, `!`, `$`, `&`), and digits.
- Each password is exactly four characters long.
- Each password has exactly one uppercase letter.
- Each password has exactly one lowercase letter.

How many possible passwords are there?

As in Example 5.12, we first choose the format of our password then count the number of passwords which conform to the chosen format. Here there are several possible formats, so let's first count the number of formats.

Each password format belongs to one of three possible "families" of formats:

1. One uppercase letter, one symbol, one lowercase letter, one digit.
2. One uppercase letter, one symbol, two lowercase letters.
3. One uppercase letter, one symbol, two digits.

Consider one possible member of each family, where we use `U` as a placeholder for the uppercase letter, `S` as a placeholder for the symbol, `L` as a placeholder for the lowercase letter, and `D` as a plceholder for the digit. So, our representative formats from each family might be `USLD`, `USLL`, and `USDD`.

Now let's count the number of formats from each family by counting the number of ways to rearrange the letters in our representative format. (This suffices to count the number of formats in the family since each format in a family is just a rearrangement of another format in the same family.)

The number of `USLD` formats is 4!; the number of `USLL` formats is $\frac{4!}{2!}$; and the number of `USDD` formats is $\frac{4!}{2!}$. Once we choose a format, we can easily count the number of passwords conforming to that format.

For a `USLD` format there are $26 \cdot 5 \cdot 26 \cdot 10$ passwords; for a `USLL` format there are $26 \cdot 5 \cdot 26^2$ passwords; and for a `USDD` format there are $26 \cdot 5 \cdot 10^2$ passwords.

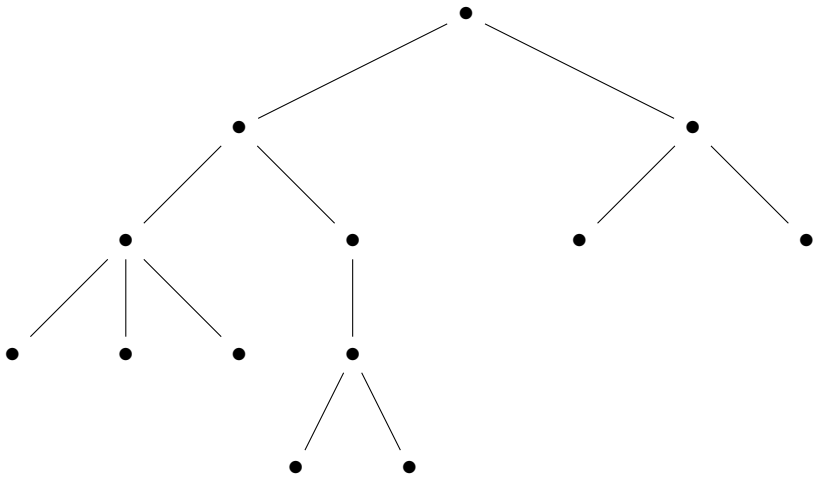
Multiplying the number of passwords in each format by the number of similar formats in the same family, then adding up all of the possibilities, the number of passwords is

$$4! \cdot 26^2 \cdot 10 \cdot 5 + \frac{4!}{2!} 26^3 \cdot 5 + \frac{4!}{2!} 26 \cdot 10^2 \cdot 5 = 2,021,760.$$

## 5.4 Trees

Sometimes it can be helpful to visualize the number of possible outcomes to an experiment by drawing a "tree" where we have one branch of the tree for each possible choice. Here a tree refers to a particular type of diagram where we have one special vertex called a ***root***, and attached to this root we have several line segments called ***branches***. At the other end of each branch we have another vertex, called a ***child*** of the root, and from that we may have more branches down to more children. If a particular child does not have any more branches below it, we call it a ***leaf***.

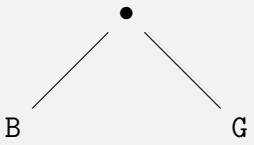
Usually when drawing trees we draw them with the root at the top, its children below it, then any children of the children below those, and so on. For example, the diagram below is a tree.



Trees are an extremely common and convenient way of organizing hierarchical information that come up in many branches of mathematics and computer science. Right now what we're trying to do is use trees to count and compute probabilities in a situation where we have to make several consecutive choices.

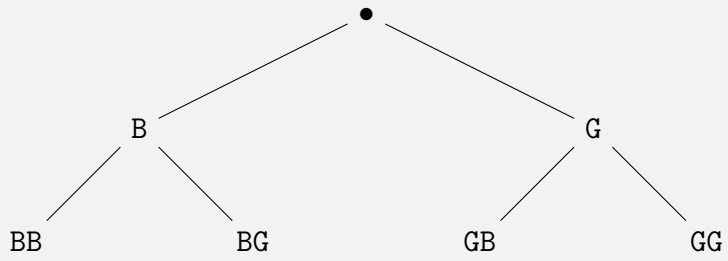
**Example 5.14.**
Suppose that in a large classroom children play a game where the children are divided into groups of three, and the teams are chosen as follows. First, a random *team captain* is chosen, and this team captain is either a boy or a girl. The team captain then chooses the second member of the team, who is again a boy or a girl. And finally, the second member of the team chooses the third member of the team, who is either a boy or a girl. Create a tree showing the possible ways a team could be selected.

We'll explain how to construct the tree one level at a time. First, the root has two children corresponding to the choices of a boy or a girl for the team captain, which we'll label as `B` or `G`.

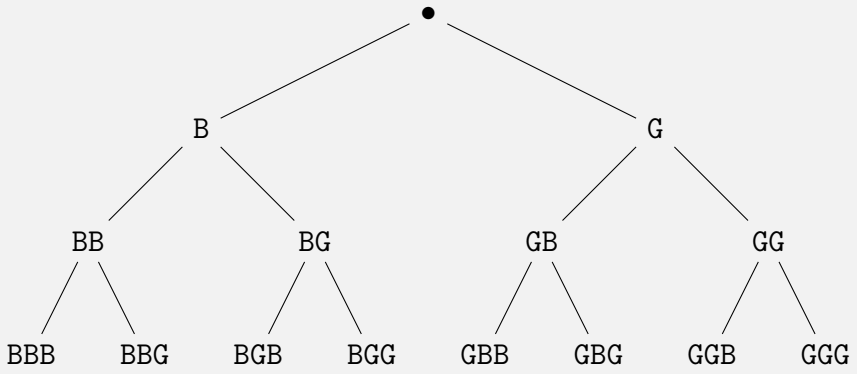



Now the team captain chooses the second member of the team, who could be a boy or a girl. Here we will label children of `B` and `G` with

two-letter strings indicating the members of the team thus far. For example, the children of B (a team captain whose a boy) will be BB and BG indicating whether the team captain (the first B for boy) chose another boy (the second B) or a girl (the G) for the second team member.




Finally, each of these nodes has two children for the third member of the team which the second member of the team gets to choose.




In the example above, the leaves of the tree (the three-letter strings at the bottom of the tree) are what we care about. The leaves of the tree above show us the all of the possible ways we could choose a three-person team of school children according to whether they are a boy or a girl, where the position of a B or G indicates the order in which a boy/girl was chosen.

We can also use trees to help us compute probabilities. Thinking of the branches of the tree as representing subsequent choices we make in going from the root to a leaf (for instance, each time we choose a boy or girl as the next member of the time), we might decide that choice will be made randomly, but with some known probability. We can label the corresponding edge of the tree with that probability, and then to compute the probability we would land at a given leaf, we multiply the probabilities

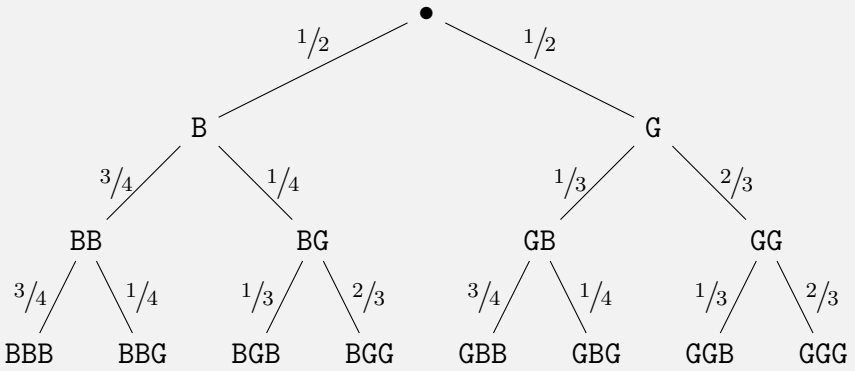
along the branches connecting the root to that leaf.

This is a wordy description of a simple idea which will probably make more sense after seeing an example.

**Example 5.15.**
Suppose, as in Example 5.14 we build a tree describing all three-student teams of boys or girls where each member of the team selects the next member. Suppose, though, that boys are more likely to select boys as the next team member, and similarly girls are more likely to select girls. In particular, suppose that the probability a boy chooses the next team member to be a boy is $3/4$, and the probability he chooses a girl is $1/4$. Suppose that girls will select another girl as the next team member with probability $2/3$ and will select a boy with probability $1/3$. Suppose the teacher chooses team captains for the teams with probability $1/2$ for both boys and girls.

What is the probability a randomly selected team will consist of a girl team captain, a boy as the second team member, and a girl as the third team member?

We simply take our tree from before, but put a label on each edge, the label telling us the probability of making each choice. According to the rules above, this tree would be as follows:




We're trying to find the probability the randomly selected team is `GBG`. The tree tells us the probability the team captain is a girl is $1/2$; this girl will select a boy as the second team member with probability $1/3$ (this is the edge labelled $1/3$ from `G` to `GB`); and the boy will choose a girl for the third team member with probability $1/4$ (this corresponds to the edge between `GB` and `GBG` which has label $1/4$). Multiplying these

values together, we see the probability of the desired team, GBG, is

$$\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{24} \approx 0.04$$

There are a few things to notice about the probabilities computed by multiplying values along the branches of the tree described in the last example. Perhaps the most obvious question is why does this work? Thus far we've said this "rule" will give the correct probabilities without justifying why that's the case (e.g., why multiply the values instead of add them or divide them or do something else?). This will be easiest to explain after we discuss conditional probability in the next section, so for now we're going to take this on faith, but we will come back and justify it later.

The second thing to point out is that there are eight possible teams, but not all teams will occur with the same probability. For example, the probability of an all boy team, BBB, is

$$\frac{1}{2} \cdot \frac{3}{4} \cdot \frac{3}{4} = \frac{9}{32} \approx 0.28$$

whereas the probability of an all girl team, GGG, is

$$\frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{4}{18} = \frac{2}{9} \approx 0.22.$$

This is our first example where we have a finite sample space, but the probability of any given event is not simply the number of elements in the event divided by the number of elements in the sample space. We'll see many other examples like this throughout the course, but since this is the first time we've seen such a thing, it's worthwhile to point it out.

**Example 5.16.**
With the same setup as Example 5.15, what is the probability a randomly selected team contains at least two girls?

Let's let $E$ denote the event where we construct a team containing at least two girls. Looking at the leaves of the tree above, we see that $E$ consists of the following teams:

$$E = \{\texttt{BGG}, \texttt{GBG}, \texttt{GGB}, \texttt{GGG}\}.$$

We can write $E$ as the disjoint union of simple events,

$$E = \{\texttt{BGG}\} \cup \{\texttt{GBG}\} \cup \{\texttt{GGB}\} \cup \{\texttt{GGG}\},$$

then the probability of $E$ is the sum of the probabilities of the simple events,

$$\Pr(E) = \Pr(\{\texttt{BGG}\}) + \Pr(\{\texttt{GBG}\}) + \Pr(\{\texttt{GGB}\}) + \Pr(\{\texttt{GGG}\}).$$

Each of these probabilities we can compute by multiplying the values indicated by the tree:

$$\begin{aligned}
\Pr(\{\texttt{BGG}\}) &= \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{2}{3} \frac{1}{12} \\
\Pr(\{\texttt{GBG}\}) &= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{24} \\
\Pr(\{\texttt{GGB}\}) &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{9} \\
\Pr(\{\texttt{GGG}\}) &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{2}{9}
\end{aligned}$$

Adding all of these together we have that the probability a randomly constructed team has at least two girls is

$$\Pr(E) = \frac{1}{12} + \frac{1}{24} + \frac{1}{9} + \frac{2}{9} = 11/24 \approx 0.46$$

**Exercise 5.6.**
Using the same setup as Example 5.15, answer the following two questions.

(a) What is the probability a randomly selected team has at least one boy and at least one girl?

(b) What is the probability a randomly selected team has at least one boy?

## 5.5 Practice Problems

**Problem 5.1.**
Suppose a bag contains eleven wooden tiles which have letters etched on them, one tile each for each of the letters `A`, `D`, `E`, `F`, `G`, `I`, `L`, `N`, `O`, `R`, `S`. If you pull the tiles out of the bag randomly one at a time and lay them across a table from left-to-right as you draw them, what is the probability you wind up spelling the word `DRAGONFLIES`?

**Problem 5.2.**
Suppose a bag contains eleven wooden tiles with letters etched into them where there are two tiles with `A`, two tiles with `M`, two tiles with `T`, and one tile each for the letters `H`, `E`, `I`, `C`, and `S`.

If you draw the eleven tiles out of the bag one at a time and lay them out across the table from left-to-right as you draw them, what is the probability that you spell out `MATHEMATICS`?

**Problem 5.3.**
Imagine that you won a set of four movie tickets, so you decided to take three of your friends with you to the movies this weekend, and suppose you have six friends that you are considering taking: Alice, Bob, Claire, Daniel, Erica, and Fred. Since you have a hard time picking who to take, you decide to put the six friends' names in a hat and randomly draw three.

(a) What is the probability that Claire is one of the three friends you select?

(b) What is the probability that both Erica and Fred are selected?

(c) What is the probability that all three selected friends have the same gender? (Alice, Claire and Erica are girls; Bob, Daniel, and Fred are boys.)

**Problem 5.4.**
Imagine an urn contains sixteen marbles, and of these eight are blue, six are green, and two are red. If you reach into the urn and pull out three marbles, one after the other and without replacing the marbles you've pulled out, what is the probability that you grabbed at least two marbles of the same color?

**Problem 5.5.**
In a five card poker hand, a *full house* occurs when you have three cards of one rank, and two cards of another rank. For example, three Jacks and two Aces is a full house.

If you were to be dealt five random cards from a shuffled deck of 52 standard playing cards, what is the probability you would get a full house?

**Problem 5.6.**
Suppose that one morning while getting ready for class you are in a hurry, not paying attention to what you're doing, and simply reach into your sock drawer and pull out two random socks. Supposing your sock drawer has twelve white socks, six black socks, four brown socks, and four blue socks. (These are individual socks, not pairs.)

What is the probability both socks are the same color?

**Problem 5.7.**
Twelve students have iPhones, and eight students have Android phones. Three students are selected at random. Find the probability that exactly one student has an Android phone and not all students have the same type of phone.

# 6 *Conditional Probability*

> *C'est une très certaine vérité que, lorsqu'il n'est pas en notre pouvoir de discerner les opinions les plus vraies, nous devons suivre le plus probable.*
> It is a very certain truth that when it is not in our power to discern the most true opinions, we must follow the most probable.
>
> René Descartes
> *Discours de la Méthode*

Sometimes we may have partial information about the outcome of an experiment and we can use ths information to help us compute probabilities.

## 6.1 Motivating example: Texas Hold 'Em

For example, suppose you have already been dealt two cards from a five-card poker hand; say you received $J\heartsuit$ and $7\diamondsuit$. What is the probability that, after receiving the other three cards, you will have a four-of-a-kind?

**Remark.**
There are poker games where you have this type of information, by the way. In *Texas Hold 'Em* each player is first dealt two cards, and then in subsequent rounds more cards are added to the middle of the table for all players to use. Players then use their two private cards (which only they know) together with the cards in the center of the table to build the best possible hand. However, betting starts after players receive their first two cards, but before the cards in the middle of the table have been revealed. Thus the players have partial information about what type of hand they might be able to build, and might like to use this information to determine the probabilities of various types of poker hands.

There are fifty cards left in the deck (we've already received two cards), and we're going to receive three more of those cards. So, given our two

initial cards, $J\heartsuit$ and $7\diamondsuit$, there are $\binom{50}{3}$ possible hands we can build. Of these, how many will result in us having a four-of-a-kind? If we're able to build a four-of-a-kind, that means we'll have four of one of these cards. That is, our only options for a four-of-a-kind are

$$(7\diamondsuit, J\heartsuit, J\clubsuit, J\diamondsuit, J\spadesuit) \text{ and } (J\heartsuit, 7\heartsuit, 7\clubsuit, 7\diamondsuit, 7\spadesuit).$$

There are only two of the possible $\binom{50}{3}$ hands, given we already have $7\diamondsuit$ and $J\heartsuit$, that result in a four-of-a-kind. Thus the probability of us being able to build a four-of-a-kind is

$$\frac{\binom{2}{1}}{\binom{50}{3}} = \frac{2}{\binom{50}{3}} \approx 0.000102.$$

**Remark.**
Notice the probability calculated above is less than half the probability for getting a four-of-a-kind we calculated in Example 5.8. The partial information we had really helped us winnow down the probability we could get a four-of-a-kind.

Let's now make a simple observation about the probability we calculated above. If we multiply and divide the fraction above by $1/\binom{52}{5}$ we would be multiplying by one (since we multiply and divide by the same thing), which doesn't change the value of the fraction. I.e.,

$$\frac{\binom{2}{1}}{\binom{50}{3}} = \frac{\binom{2}{1}/\binom{52}{5}}{\binom{50}{3}/\binom{52}{5}}.$$

Now notice the numerator $\binom{2}{1}/\binom{52}{5}$ is the probability of a five-card poker hand which is a four-of-a-kind containing both $J\heartsuit$ and $7\diamondsuit$. (This is because, as we noted above, there are only two five-card hands which are four-of-a-kind and contain both $J\heartsuit$ and $7\diamondsuit$). The denominator, $\binom{50}{3}/\binom{52}{5}$ is the probability of five-card poker hand containing $J\heartsuit$ and $7\diamondsuit$, regardless of what the other cards are. (This is because we've already selected these two cards and just need to select the other three from the remaining fifty.)

If we let $E$ be the event we get a four-of-a-kind and $F$ the event that our five-card hand contains $J\heartsuit$ and $7\diamondsuit$, notice that $E \cap F$ is the event we have a four-of-a-kind which contains $J\heartsuit$ and $7\diamondsuit$. Thus the $\binom{2}{1}/\binom{52}{5}$ in

the numerator is $\Pr(E \cap F)$, and the $\binom{50}{3}/\binom{52}{5}$ is $\Pr(F)$. So the probability that we get a four-of-a-kind given that we already have $7\diamondsuit$ and $J\heartsuit$ can be written as

$$\frac{\Pr(E \cap F)}{\Pr(F)}.$$

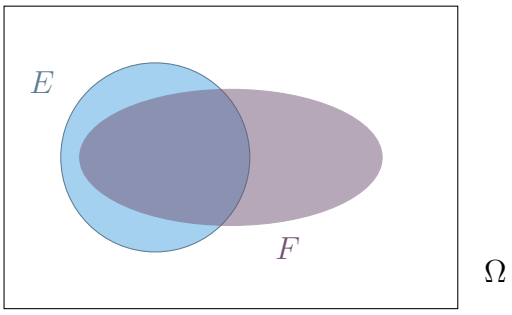
## 6.2 Definition of conditional probability

These types of calculations are extremely common in probability theory, and the discussion above motivates the next definition. In general, the probability an event $E$ takes place given that we already know the event $F$ will take place is called the ***conditional probability of $E$ given $F$*** which is denoted $\Pr(E|F)$. We can calculate this probability with the same formula as in the example above:

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

You should interpret these conditional probabilities as being the probability of some sort of "subexperiment" within our original experiment. Let's draw a Venn diagram to better understand what's happening.

Imagine that the sample space $\Omega$ of our experiment is a rectangle, and $E$ and $F$ are two events that live inside of $\Omega$:

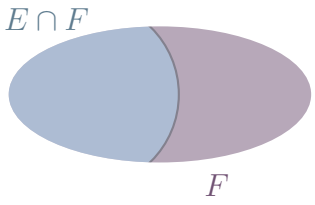



When we say that $F$ is given, that means we know that the outcome of our experiment will live in $F$. We may still not know exactly which point in $F$ it is, but it will be some element of $F$. So, we can consider the "subexperiment" where the original sample space is replaced by $F$, and the events are subsets of $F$ instead of arbitrary subsets of the original $\Omega$.

If we replace our original experiment with one where the outcomes are constrained to $F$, then we also need to change the way we calculate probabilities. Our Pr function won't work anymore because the probability of our sample space (which is now $F$) won't be 1, it'll be $\Pr(F)$. To fix this,

let's just divide everything by $\Pr(F)$: this will make it so the probability of the entire sample space (now $F$ in our "subexperiment") is 1. This gives us a new probability function for the events inside of $F$ which is denoted $\Pr(-|F)$ where the dash means this is where we'd write the event inside of $F$ whose probability we want to calculate. For example, the probability of the event $G \subseteq F$ is $\Pr(G|F)$ and is calculated by $\Pr(G|F) = \frac{\Pr(G)}{\Pr(F)}$ because of the scaling described above.

Now, if $E$ an event that is not contained in $F$ and we're given that $F$ is going to occur, the only option if $E$ occurs as well is that the outcome of the experiment is in the overlap between $E$ and $F$ which is precisely $E \cap F$.




As another example to explain the idea, imagine that you've been practcing at playing darts for the last several weeks and have gotten to be very good at getting the dart to land more-or-less where you're intending it to land. Your skill at darts is the "partial information" that we can use to update our probabilities. Recall that earlier when we discussed throwing darts, we treated the point where the dart landed as a random point on the board, and so the probability you hit the bullseye on the board, for instance, was

$$\Pr(\text{Bullseye}) = \frac{\text{Area(Bullseye)}}{\text{Area(Dart Board)}}.$$

But given that you've practiced enough that you can, say, always land in the middle third of the board, the probability you hit the bullseye becomes

$$\frac{\text{Area(Portion of Bullseye in Middle Third of Board)}}{\text{Area(Middle Third of Board)}}.$$

Multiplying and dividing the numerator and denominator by the area of the entire board, we can write this as

$$\frac{\text{Area(Portion of Bullseye in Middle Third of Board)}/\text{Area(Board)}}{\text{Area(Middle Third of Board)}/\text{Area(Board)}}.$$

This is equal to

$$\frac{\Pr(\text{Bullseye in Middle Third of Board})}{\Pr(\text{Middle Third of Board})}$$

which in terms of conditional probability is

$$\Pr(\text{Bullseye} \mid \text{Middle Third of Board}).$$

**Remark.**
If this idea of conditional probability seems a little bit strange right now, don't worry too much about it: it will eventually start to make sense as we see more examples and as you do some calculations on your own.

## 6.3 Examples

Let's consider a few concrete examples to help solidify the idea of conditional probability.

**Example 6.1.**
Suppose a congressional of subcomittee is randomly selected from a committee which already contains six Republicans and four Democrats. If the sub committee consists of three members and is required to be bipartisan (i.e., must consist of both Republicans and Democrats), what is the probability the committee contains two Democrats?

Here the "partial information" we are given is that the committee must be bipartisan. Let $B$ denote the event a randomly selected subcommittee is bipartisan, and let $D$ be the event the subcommittee contains two Democrats. Then the probability the submittee contains two Democrats given that it is bipartisan is

$$\Pr(D|B) = \frac{\Pr(D \cap B)}{\Pr(B)}$$

Let's first think about $\Pr(B)$. If the subcommittee *is not* bipartisan, then that means it contains all Republicans or all Democrats, and it might be easier to find $\Pr(B^c)$ than to find $\Pr(B)$ directly.

If we had a non-bipartisan subcommittee, we would have to choose our three Republican members or our three Democratic members. Di-

viding by all the ways we can choose three members from the 10-member committee we have

$$\Pr(B^c) = \frac{\binom{6}{3} + \binom{4}{3}}{\binom{10}{3}}$$

and so

$$\Pr(B) = 1 - \Pr(B^c) = 1 - \frac{\binom{6}{3} + \binom{4}{3}}{\binom{10}{3}}$$

As for the numerator, $\Pr(D \cap B)$, we want the probability of containing two Democratic members and being bipartisan. Since the committee consists of only three members, this means there are two Democrats and one Republican. So let's simply choose the two Democrats and the one Republican, divided by all the ways to choose a 3-member subcommittee:

$$\Pr(D \cap B) = \frac{\binom{4}{2} \cdot \binom{6}{1}}{\binom{10}{3}}.$$

Putting all fo this together, and simplifying a little, we have

$$\Pr(D|B) = \frac{\binom{4}{2} \cdot \binom{6}{1}}{\binom{10}{3} - \binom{6}{3} - \binom{4}{3}} = \frac{36}{96} = 0.375.$$

**Example 6.2.**
Suppose a family has two children. What is the probability the family has two girls if the oldest child is a girl?

Let's write the possible outcomes of our experiment as two-character strings where each character is `B` or `G`, indicating if a child is a boy or a girl, where the gender of the oldest child is the character on the left, and th gender of the youngest child is the character on the right. E.g., `BG` would mean the oldest child is a boy and the youngest child is a girl.

The sample space of this experiment is then

$$\{\texttt{BB}, \texttt{BG}, \texttt{GB}, \texttt{GG}\}.$$

Now let $T$ be the event the family has two girls, and $G$ the event the

oldest child is a girl. So, $T = \{\texttt{GG}\}$ and $G = \{\texttt{GB}, \texttt{GG}\}$. Notice that, assuming all possibilities are equally likely,

$$\Pr(T) = 1/4, \quad \Pr(G) = 1/2.$$

Now we compute

$$\Pr(T|G) = \frac{\Pr(T \cap G)}{\Pr(G)} = \frac{\Pr(\{\texttt{GG}\})}{\Pr(G)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

The solution to the previous example should seem very intuitive, but let's modify the example just a little bit and we'll see that we get a very counter-intuitive result.

**Example 6.3.**
Suppose a family has two children. If we ask one of the parents "Do you have a daughter?" and they simply answer "Yes.", then what is the probability both children are girls?

As in Example 6.2 the sample space still consists of four elements, and we'll let $T$ be the event the family has two girls. However, now $G$ is the event the family has at least one girl: we only know they have a girl, not whether she is the oldest child or not. So, $G$ is the event

$$G = \{\texttt{BG}, \texttt{GB}, \texttt{GG}\}.$$

Now when we compute the probability both children are girls we have

$$\Pr(T|G) = \frac{\Pr(T \cap G)}{\Pr(G)} = \frac{\Pr(\{\texttt{GG}\})}{\Pr(G)} = \frac{1/4}{3/4} = \frac{1}{3}.$$

Notice that even though Example 6.2 and Example 6.3 seem almost like identical experiments, they have different answers. This is entirely because we have different partial information in the two problems.

## 6.4 Consequences of the definition

Let's now make some simple observations about our definition of conditional probability and see some useful consequences of the definition.

We defined the conditional probability of $E$ given $F$ as

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

Notice that if we multiply both sides of this equation by $\Pr(F)$ we have

$$\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F).$$

That is, we can use conditional probability to give us a formula for probabilities of intersections. This is sometimes useful because you may already know the conditional probability from the context of a problem and want to use that information to calculate other probabilities.

**Exercise 6.1.**
Show that we can also write $\Pr(E \cap F)$ as $\Pr(F|E) \cdot \Pr(F)$ by a simple manipulation of the above equation.

**Example 6.4.**
Suppose a student applying to their preferred college has an 80% chance of being accepted, and that 60% of students at this college live on campus. What is the probability the student both gets accepted to the school and then also lives on campus?

Let $A$ be the event the student is accepted, and $L$ the event they live on campus. We're already told $\Pr(L|A) = 0.6$, but what we want is $\Pr(A \cap L)$. Using the formula above (and Exercise 6.1 which tells us how to swap the order of the events), we have

$$\Pr(A \cap L) = \Pr(L|A) \cdot \Pr(A) = 0.6 \cdot 0.8 = 0.48,$$

and so there is a 48% chance the student will both get accepted to the school and live on campus.

**Exercise 6.2.**

(a) Show that for any two events $E$ and $F$, $\Pr(E^c|F) = 1 - \Pr(E|F)$.

(b) Is it true that $\Pr(E|F) = 1 - \Pr(E|F^c)$?

**Example 6.5.**
Suppose a certain high school has 400 students, 120 of which are boy sand the remaining 280 are girls. Suppose that of the 120 boys, 72 (60%) are enrolled in a math course; and of the 280 girls, 224 (80%) are encrolled in a math course.

(a) What is the probability a randomly selected student is a girl enrolled in a math course?

(b) What is the probability a randomly selected student is a boy who is not enrolled in a math course?

Let's let $G$ be the event a randomly selected student is a girl, $B$ the event a randomly selected student is a boy, and $M$ the event a randomly selected student is enrolled in a math course. We are told that

$$\Pr(B) = \frac{120}{400} = 0.3$$
$$\Pr(G) = \frac{280}{400} = 0.7$$
$$\Pr(M|G) = \frac{224}{280} = 0.8$$
$$\Pr(M|B) = \frac{72}{120} = 0.6$$

(a) We are interested in the event $M \cap G$. Using the expression above we compute

$$\Pr(M \cap G) = \Pr(M|G) \cdot \Pr(G) = 0.8 \cdot 0.7 = 0.56$$

so there is a 56% chance a randomly selected student is a girl enrolled in a math calss.

(b) We are interested in $M^c \cap B$. We are told $\Pr(M|B) = 0.6$, and by part (a) of Exercise 6.2 we know $\Pr(M^c|B) = 0.4$. Now we can compute

$$\Pr(M^c \cap B) = \Pr(M^c|B) \cdot \Pr(B) = 0.4 \cdot 0.3 = 0.12,$$

and so there is only a 12% chance a randomly selected student is a boy who is not enrolled in a math class.

Let's notice our observation that

$$\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F)$$

can be extended to intersections with more than two events. For example, by the definition of conditional probability we know

$$\Pr(E|F \cap G) = \frac{\Pr(E \cap F \cap G)}{\Pr(F \cap G)}$$

which we can rewrite as

$$\Pr(E \cap F \cap G) = \Pr(E|F \cap G) \cdot \Pr(F \cap G).$$

We can then rewrite $\Pr(F \cap G)$ as $\Pr(F|G) \cdot \Pr(G)$ to turn the above into

$$\Pr(E \cap F \cap G) = \Pr(E|F \cap G) \cdot \Pr(F|G) \cdot \Pr(G).$$

**Remark.**
Since $E \cap F \cap G = F \cap E \cap G = F \cap G \cap E = \cdots$, we can actually

rewrite the probability of $E \cap F \cap G$ in many different ways:

$$\begin{aligned}\Pr(E \cap F \cap G) &= \Pr(E|F \cap G) \cdot \Pr(F|G) \cdot \Pr(G) \\ &= \Pr(F|E \cap G) \cdot \Pr(E|G) \cdot \Pr(G) \\ &= \Pr(F|E \cap G) \cdot \Pr(G|E) \cdot \Pr(E) \\ &= \Pr(G|E \cap F) \cdot \Pr(E|F) \cdot \Pr(F) \\ &\vdots\end{aligned}$$

In general, we have the following "chain rule" for probabilities intersections:

**Proposition 6.1.**
*For any events $E_1$, $E_2$, $E_3$, ..., $E_n$ in a sample space $\Omega$ we may rewrite the probability of the intersection of all of the events as a product of conditional probabilities using the following formula:*

$$\begin{aligned}&\Pr(E_1 \cap E_2 \cap \cdots \cap E_n) \\ =&\Pr(E_1|E_2 \cap \cdots \cap E_n) \cdot \Pr(E_2|E_3 \cap \cdots \cap E_n) \cdot \ldots \cdot \Pr(E_{n-1}|E_n) \cdot \Pr(E_n)\end{aligned}$$

*Proof.*
We will prove this proposition with a proof technique called induction where we prove the simplest possible scenario first, and then show how the more complicated scenarios can be expressed in terms of simpler ones.

The base case (simplest scenario) here is the case when $n = 2$. In this case the proposition claims

$$\Pr(E_1 \cap E_2) = \Pr(E_1|E_2) \cdot \Pr(E_2)$$

which we already know is true from the definition of conditional probability.

We now assume the proposition has been proven for $n - 1$ events,

in which case the proposition says

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_{n-1}) = \Pr(E_1|E_2 \cap \cdots \cap E_{n-1}) \cdot \ldots \cdot \Pr(E_{n-2}|E_{n-1}) \cdot \Pr(E_n).$$

Supposing this is true, we can now easily extend the formula for $n$ events by rewriting this as an intersection of two events, then applying the base case and the induction hypothesis:

$$\begin{aligned} &\Pr(E_1 \cap E_2 \cap \cdots \cap E_n) \\ =&\Pr(E_1 \cap [E_2 \cap \cdots \cap E_n]) \\ =&\Pr(E_1|E_2 \cap \cdots \cap E_n) \cdot \Pr(E_2 \cap \cdots \cap E_n) \\ =&\Pr(E_1|E_2 \cap \cdots \cap E_n) \cdot \Pr(E_2|E_3 \cap \cdots \cap E_n) \cdot \ldots \cdot \Pr(E_{n-1}|E_n) \cdot \Pr(E_n). \end{aligned}$$

□

**Example 6.6.**
Suppose an urn contains thirty marbles. Of these, seven are blue and twenty-three are red. If three marbles are randomly selected from the bag without replacement, what is the probability all three marbles are red?

Let $E_i$ denote the event the $i$-th marble is red. Then the event all three are red is $E_3 \cap E_2 \cap E_1$. Using our proposition above for the probability of intersections we can write this as

$$\Pr(E_3 \cap E_2 \cap E_1) = \Pr(E_3|E_2 \cap E_1) \cdot \Pr(E_2|E_1) \cdot \Pr(E_1).$$

Notice $\Pr(E_1) = \frac{23}{30}$ since there are twenty-three red marbles out of the thirty marbles in the urn. Now, for $\Pr(E_2|E_1)$ we are interested in the event where the second marble is red *assuming* the first marble was red as well. Since the first marble was red, by assumption, that means that twenty-two of the remaining twenty-nine marbles are red and so $\Pr(E_2|E_1) = \frac{22}{29}$. Similarly, $\Pr(E_3|E_2 \cap E_1)$ is easy to compute because we are assuming that the first two marbles were red (this is the given partial information, the $E_2 \cap E_1$). Thus $\Pr(E_3|E_2 \cap E_1) = \frac{21}{28}$.

Multiplying everything together we have

$$\Pr(E_3 \cap E_2 \cap E_1) = \frac{21}{28} \cdot \frac{22}{29} \cdot \frac{23}{30} = \frac{10626}{24360} \approx 0.436.$$

So, there's a little more than a 43% chance of getting three red marbles.

## 6.5 The law of total probability

We now want to discuss a simple technique for computing probabilities of events in terms of "simple" events. In order to discuss this, though, we have to make a quick set-theoretic definition.

A ***(finite) partition*** of a set $\Omega$ is a finite collection of subsets $F_1$, $F_2$, ..., $F_n$ that satisfy the following two conditions:

1. The sets are *pairwise disjoint*. This means that for any two sets $F_i$ and $F_j$, where $i \neq j$ (so they are two distinct subsets from our collection), the sets are disjoint: $F_i \cap F_j = \emptyset$.

2. The union of these subsets is the entire set: $\Omega = \bigcup_{i=1}^{n} F_i$.

What this means is that we can take our set $\Omega$, and chop it up into finitely-many pieces such that the pieces don't overlap and every element of the original, big set is contained in one of the smaller pieces.

Now, if $F_1$, $F_2$, ..., $F_n$ is a partition of a sample space, then for any event $E$ we can consider the overlap of $E$ with each of the partition sets $F_1$ through $F_n$. I.e., we can consider $E \cap F_1$, $E \cap F_2$, ..., $E \cap F_n$. Notice this is actually a partition of $E$.

**Exercise 6.3.**
Show that if $F_1$, $F_2$, ..., $F_n$ is a partition of $\Omega$, then for any $E \subseteq \Omega$, the sets $E \cap F_1$, $E \cap F_2$, ..., $E \cap F_n$ form a partition of $E$.

Since the $E \cap F_i$ form a partition of $E$, we can write

$$\Pr(E) = \Pr(E \cap F_1) + \Pr(E \cap F_2) + \cdots + \Pr(E \cap F_n).$$

Writing each $\Pr(E \cap F_i)$ as $\Pr(E|F_i) \cdot \Pr(F_i)$ we have the following expression, called the ***law of total probability***:

$$\Pr(E) = \sum_{i=1}^{n} \Pr(E|F_i) \cdot \Pr(F_i)$$

The simplest examples of this come from the simplest type of partition. If $F$ is any event, then $F$ and $F^c$ form a partition of the sample space $\Omega$. For any event $E$ the law of total probability then tells us the probability of $E$ may be written as

$$\Pr(E) = \Pr(E|F) \cdot \Pr(F) + \Pr(E|F^c) \cdot \Pr(F^c).$$

**Example 6.7.**
Suppose sore throat is present in 70% of people with the flu, and is present in 15% of people without the flu. If 10% of people in a given population have the flu, what is the probability a randomly selected person has a sore throat?

To apply the law of total probability we need a partition, and one easy way to get a partition is to have two complementary events. If we let $F$ be the set of people with the flu, so $F^c$ is the set of people without the flu, then we have our partition. Letting $E$ be the set of people with a sore throat, the law of total probability tells us

$$\Pr(E) = \Pr(E|F)\Pr(F) + \Pr(E|F^c)\Pr(F^c).$$

We know each probability on the right-hand side from the problem. Since 10% of people have the flu, we know $\Pr(F) = 0.1$; this also means 90% of people don't have the flu, so $\Pr(F^c) = 0.9$. We know that of the people that have the flu, 70% have a sore throat and so $\Pr(E|F) = 0.7$; and only 15% of people without the flu have sore throat, so $\Pr(E|F^c) = 0.15$. Putting all of this together we have

$$\Pr(E) = 0.7 \cdot 0.1 + 0.15 \cdot 0.9 = 0.205$$

So there's a 20.5% chance a randomly selected person has a sore throat.

**Example 6.8.**
Imagine there are three urns containing colored marbles. The first urn contains 14 blue marbles and 6 red marbles; the second urn contains 10 blue marbles and 15 green marbles; the third urn contains 5 blue marbles, 5 green marbles, and 10 red marbles. If you randomly select one urn , and from that urn randomly select one marble, what is the probability you draw a blue marble?

Let $B$ be the event we get a blue marble; our goal is to compute $P(B)$. Letting $U_1$ denote the event where we select a marble from the first urn, $U_2$ the event where we select a marble from the second urn, and $U_3$ the event where we select a marble from the third urn, the three events $U_1$, $U_2$, and $U_3$ form a partition of the sample space. By the law of total probability we can then compute $P(B)$ as follows, assuming the probability of choosing each urn is equally likely and the probabilitiy of drawing a blue marble from a given urn using the proportion of blue marbles described above we have

$$\begin{aligned} P(B) &= P(B|U_1)P(U_1) + P(B|U_2)P(U_2) + P(B|U_3)P(U_3) \\ &= \frac{14}{20} \cdot \frac{1}{3} + \frac{10}{25} \cdot \frac{1}{3} + \frac{5}{20} \cdot \frac{1}{3} \\ &= \frac{9}{20} = 0.45 \end{aligned}$$

## 6.6 Bayes' formula

In Example 6.7 we wanted to know the probability a randomly selected person had a sore throat. That was a silly question because it's not really something anyone cares about, so let's modify that problem to answer a more interesting question. Using the same values as before (10% of the population has the flu; 70% of people with the flu have a sore throat; and 15% of people without the flu have a sore throat), consider the following: If someone has a sore throat, what is the probability they have the flu?

**Remark.**
This is a question that someone might genuinely want to know the

answer to, and it is also indicative of a more general type of real-world question. If a given symptom can be caused by several different diseases, you might want to know the likelihood someone has a particular disease if they have that symptom. For instance, if the symptom is severe headaches, the causes might be allergies, stress, caffeine withdrawal, or brain cancer. A doctor may be very interested in knowing which of these causes is the most likely as that may influence the type of treatment they recommend.

Let's again let $F$ be the event the person has the flu, and $E$ the event they have a sore throat. Earlier we used the law of total probability to compute $\Pr(E)$, but now we want to compute $\Pr(F|E)$: the probability someone has the flu, given they have a sore throat.

By the definition of conditional probability we know

$$\Pr(F|E) = \frac{\Pr(F \cap E)}{\Pr(E)}.$$

We do not know what $\Pr(F \cap E)$ is directly from the information we are given, but we can compute it as

$$\Pr(F \cap E) = \Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F).$$

We also are able to compute $\Pr(E)$ using the law of total probability:

$$\Pr(E) = \Pr(E|F) \cdot \Pr(F) + \Pr(E|F^c) \cdot \Pr(F^c).$$

Plugging this into the right-hand side of our expression for $\Pr(F|E)$ above we have

$$\Pr(F|E) = \frac{\Pr(E|F) \cdot \Pr(F)}{\Pr(E|F) \cdot \Pr(F) + \Pr(E|F^c) \cdot \Pr(F^c)}.$$

And if we plug in the values for $\Pr(E|F)$, $\Pr(F)$, etc. that we have from before this is

$$\begin{aligned}
\Pr(F|E) &= \frac{\Pr(E|F) \cdot \Pr(F)}{\Pr(E|F) \cdot \Pr(F) + \Pr(E|F^c) \cdot \Pr(F^c)} \\
&= \frac{0.7 \cdot 0.1}{0.7 \cdot 0.1 + 0.15 \cdot 0.9} \\
&= \frac{0.07}{0.205} \\
&\approx 0.34146.
\end{aligned}$$

Thus there is about a 34% chance that someone with the sore throat has the flu.

The calculation above is an example of a more general formula called ***Bayes' formula***, and was first described by the 18th century English statistician Thomas Bayes. (Interestingly, Bayes never actually published Bayes formula. The formula was found in unpublished notes of Bayes by Richard Price sometime after Bayes' death.)

In general, Bayes' says that if $F_1$, $F_2$, ..., $F_n$ is a partition of a sample space $\Omega$, then for any event $E \subseteq \Omega$, the probability of $F_i$ given $E$ can be compute as

$$\begin{aligned}\Pr(F_i|E) &= \frac{\Pr(F_i \cap E)}{\Pr(E)} \\ &= \frac{\Pr(E|F_i) \cdot \Pr(F_i)}{\Pr(E)} \\ &= \frac{\Pr(E|F_i) \cdot \Pr(F_i)}{\sum_{j=1}^{n} \Pr(E|F_j) \cdot \Pr(F_j)}\end{aligned}$$

The first line above is simply the definition of conditional probability; the second line is our "trick" for rewriting the probability of an intersection in terms of conditional probability; and the last line is applying the law of total probability to compute the denominator.

Bayes' formula has numerous real-world applications, a few of which we will touch on in the examples below. But first let's consider an example where we can use Bayes' formula to determine the likelihood a patient has a disease given that a test for the disease came back positive.

**Example 6.9.**
Suppose that you have been feeling very ill lately and so you decide to go to the doctor. The doctor determines from your symptoms that there's a chance you may have very rare, but serious illness. The disease in question affects only 1 in every 1000 people, but because the disease very serious if you do have it, the doctor decides to run a blood test. No test is completely perfect, but this particular test is known to be reasonably accurate: it correctly identifies someone with the disease as having the disease 99% of the time, and gives a false positive (saying you have the disease when you don't) only 1% of the time.

Unfortunately, the blood test comes back positive. Given the accuracy of the test described above, what is the probability that you actually have the disease?

Let's let $D$ denote the event that you really do have the disease, and $T$ the event that the test comes back positive. Our goal is to compute $\Pr(D|T)$ and we can use Bayes' formula for this. From the information given about the rarity of the disease and the accuracy of the test known the following:

$$\begin{aligned}\Pr(T|D) &= 0.99\\ \Pr(T|D^c) &= 0.01\\ \Pr(D) &= 0.001\\ \Pr(D^c) &= 0.999\end{aligned}$$

(The last two probabilities come from the fact the disease affects only $1/1000$ of the population.)

Applying Bayes' formula we compute

$$\begin{aligned}\Pr(D|T) &= \frac{\Pr(T|D)\cdot\Pr(D)}{\Pr(T|D)\cdot\Pr(D)+\Pr(T|D^c)\cdot\Pr(D^c)}\\ &= \frac{0.99\cdot 0.001}{0.99\cdot 0.001+0.01\cdot 0.999}\\ &= \frac{0.00099}{0.00099+0.00999}\\ &= \frac{0.000999}{0.01098}\\ &= 0.0901639\end{aligned}$$

That is, there is about a 9% chance you actually have the disease given that the test came back positive.

The first time you see an example like the one above, you're probably very surprised by the low probability. If the test is 99% accurate, shouldn't that probability be more like 99% instead of 9%? This is a common misconception, so let's take the time to slowly think through what's happening.

If we had a group of 1000 people, we would expect only 1 of them to actually have the disease. However, the blood test is known to give false positives 1% of the time. So, if we gave the blood test to everyone in our group of 1000% people, it would probably correctly say that one person had the disease (since the test does this 99% of the time), but it would also say

that 1% of those remaining 999 people without the disease actually have it. Now, 1% of 999 is 9.99, and let's round that up to 10 just to have nice numbers. So, our test would tell 11 people they have the disease (1 person with the disease, and then the 10 false positives). That is, if you were one of the 11 people that had a the positive blood test, there's only a $1/11 \approx 0.091$ (or about 9%) chance you're actually the person with the disease.

**Example 6.10.**
Continuing with Example 6.9 and the ensuing discussion, suppose you're a mathematically enlightened person and so you realize the chance you have the disease given a positive blood test isn't actually that high and so you go get a second opinion, and another doctor runs the same test again and it comes back positive. What is the probability you really have the disease?

In our earlier calculation we had assumed $\Pr(D) = 0.001$ since one in 1000 people have the disease. When getting a second opinion, however, we should not use the same probability. We are now part of a much smaller group of people, those who tested positive for the disease. As discussed above, for those people there's about a 9% chance of actually having the disease (being that one person with the disease in the eleven people that tested positive for the disease), so in applying Bayes' formula again we will "update" our $\Pr(D)$ value from 0.0001 to 0.09. (Notice this also chance $\Pr(D^c)$ to 0.091.) Now we compute

$$\begin{aligned}\Pr(D|T) &= \frac{\Pr(T|D)\cdot\Pr(D)}{\Pr(T|D)\cdot\Pr(D)+\Pr(T|D^c)\cdot\Pr(D^c)}\\ &= \frac{0.99\cdot 0.09}{0.99\cdot 0.09+0.01\cdot 0.91}\\ &= \frac{0.0891}{0.0982}\\ &= 0.90733\end{aligned}$$

That is, if our test comes back positive when we get a second opinion, there's about a 90% chance we really do have the disease.

**Remark.**
The examples above are shamelessly stolen from the YouTube channel Veritasium's video on Bayes' formula: https://youtu.be/R13BD8qKeTg.

**Exercise 6.4.**
Continuing the discussion from Example 6.9 and Example 6.10, suppose you insisted on a third opinion. Using the the fact that you've already had two tests come back positive, if a third test also comes back positive, what is the probability you have the disease?

In computing $\Pr(F_i|E)$ using Bayes' formula for a partition $F_1$, $F_2$, ..., $F_n$, the value of $\Pr(F_i)$ used is sometimes called the ***prior probability*** of $F_i$ and is the probability $F_i$ occurs without having any additional information. The value of $F_i$ given some additional information, i.e. $\Pr(F_i|E)$, is called the ***posterior probability*** and represents our "updated" probability calculation once we have some additional information. For instance, in the examples above the prior probability $\Pr(D)$ is the probability you have a certain disease *having no additional information that might indicate you have the disease or not*, such as the blood test. The posterior probability, $\Pr(D|T)$, is our updated calculation that you may have the disease, given that you had a positive blood test.

In many real-life applications the prior probability may not be known, and so we may need to use an "educated guess" at what this probability is. In the disease example, you either have the disease or you don't – there's not really anything random or probabilistic about it. However, since we don't know if you have the disease or not, without any additional information we may estimate the probability you have the disease by the proportion of people in the population with the diease. However, as we gain more information (e.g., positive blood tests) we are able to update the prior probability over time. The hope, of course, is that as we gain more and more information our prior probability becomes closer and closer to the true value.

Let's have another example of using Bayes' formula with a simple example from an area of computer science called *machine learning*, where we try

to train a computer to solve a problem without explicitly telling the computer how to solve it. One common family of problems in machine learning are *classification problems*, where we want to take given input data and categorize it into different types. There may be too many features of the data for a human to realisitically describe how all possible features fit together to solve the classification problem, so we might try to get the computer to figure out how those features fit together by feeding the computer data that has already been classified. When presented with new, unclassified data later, we want the computer to be able to automatically decide how to classify it.

One simple example of such a classification problem is spam filtering in email. Apps like Gmail are able to automatically determine if an incoming message is spam or not and decide whether to put the message in your inbox, or into a spam folder, and they are (usually) extremely good at discerning spam messages for legitimate emails. In the 90's, however, spam filtering wasn't nearly as good and it was not uncommon for your inbox to be completely filled with spam you didn't care about. After some researchers at Stanford and Microsoft published a paper in 1998 about how to use Bayes' formula to determine if an email was spam or not, spam filtering started to improve dramatically. At this point spam filtering is so good that it's relatively rare that spam pops up in your inbox or that legitimate emails get marked as spam. In the example below we work through a simplified version of this using Bayes formula to determine the probability a newly arrived email is spam or not.

**Example 6.11.**
Suppose we are given a list of phrases which we know are common in spam and uncommon in legitimate emails. (We might have such a list by compiling data about which emails users mark as spam if spam happens to make it into their mailbox, or which legitimate emails they remove from their spam folder.) For the sake of example, let's say we have three particularly spammy phrases,

*"You've been approved"*, *"Meet your soulmate"*, and *"Unclaimed assets"*.

Now suppose a new email has arrived, and we want to determine if the email is spam or not. We might write a line or two of code to have the computer scan through the email and see if it contains any of

our known spammy phrases. Let's let $Y$, $M$, and $U$ denote the events where the email contains the three phrases above (the first letter of the phrases telling us the "name" of the event). We now imagine the set of all emails we receive as being partitioned into two complementary pieces: $S$ is the set of spam emails, and $S^c$ the set of legitimate emails. We want to determine the probability an email is spam (the event $S$ occurs) or legitimate (the event $S^c$ occurs), given that the new message contains one of our spammy phrases.

Let's suppose that from previous data we've collected about email we know the following:

- 75% of all emails are spam, so 25% of emails are legitimate;
- 50% of spam emails contain the phrase *"You've been approved"*, and only 15% of legitimate emails contain this phrase;
- 30% of spam emails contain the phrase *"Meet your soulmate"*, but only 1% of legitimate emails contain that phrase;
- 70% of spam emails mention *"Unclaimed assets"*, compared to only 1% of legitimate emails.

That is, we know the following probabilities:

$$\begin{aligned} \Pr(S) &= 0.75 & \Pr(S^c) &= 0.25 \\ \Pr(Y|S) &= 0.5 & \Pr(Y|S^c) &= 0.15 \\ \Pr(M|S) &= 0.3 & \Pr(M|S^c) &= 0.01 \\ \Pr(U|S) &= 0.7 & \Pr(U|S^c) &= 0.01 \end{aligned}$$

If a newly arrived message contains the phrase *"Unclaimed assets"*, what's the probability that email is spam? To answer this, we want to compute $\Pr(S|U)$ and we can find this with Bayes' formula:

$$\begin{aligned} \Pr(S|U) &= \frac{\Pr(U|S) \cdot \Pr(S)}{\Pr(U|S) \cdot \Pr(S) + \Pr(U|S^c) \cdot \Pr(S^c)} \\ &= \frac{0.75 \cdot 0.75}{0.7 \cdot 0.075 + 0.01 \cdot 0.025} \\ &= \frac{0.525}{0.5275} \\ &= 0.995 \end{aligned}$$

So there's a 99.5% chance this message is spam. We may then have the computer automatically put such a message in the spam filter if the probability of it being spam is greater than, say 99%.

**Exercise 6.5.**
Continuing the discussion from Example 6.11, suppose a message contains multiple spammy phrases. From experience we may know that emails with several spammy phrases are more likely to be spam than those with a single spammy phrase, and legitimate emails very, very rarely contains multiple spammy phrases. For example, suppose know that 40% of spam conatins both phrases *"Unclaimed assets"* and *"You've been approved"*, but only 0.1% of legitimate emails contain both phrases.

If a new email contains both of these phrases, what is the probability it is spam?

## 6.7 Independence

Recall that we motivated conditional probability by saying that having some "partial information" (this is the $F$ in $\Pr(E|F)$) can be useful in computing probabilities. E.g., knowing that two of the cards in a five-card poker hand are $J\heartsuit$ and $7\diamondsuit$ helps us determine the likelihood that we will get a five-card poker hand with four-of-a-kind.

However, sometimes the partial information doesn't actually help us: sometimes knowing that event $F$ must happen doesn't change the probability the event $E$ will happen. Intuitively, not all events influence one another. For example, if you roll two distinguishable dice – say one is blue and one is red – knowing the red die rolls a three doesn't tell you anything about the blue die will roll. In situations like this, we say the two events are *independent*.

To be more precise, we say that two events $E$ and $F$ are ***independent*** if $\Pr(E|F) = \Pr(E)$. That is, the "partial knowledge" from the event $F$ tells us nothing about $E$. Before giving any concrete examples, let's make a

simple observation about this definition: it is symmetric in $E$ and $F$. That is, if $\Pr(E|F) = \Pr(E)$, then we also have $\Pr(F|E) = \Pr(F)$.

To see this is true, let's suppsoe we already know $\Pr(E|F) = \Pr(F)$ and try to show that $\Pr(F|E)$ must equal $\Pr(F)$. To do this we will write down the definition of the conditional probabilitiy $\Pr(F|E)$ and try to use some of the knowledge we've developed earlier in this chapter, together with our assumption $\Pr(E|F) = \Pr(E)$, to show this must equal $\Pr(F)$:

$$\begin{aligned}
\Pr(F|E) &= \frac{\Pr(F \cap E)}{\Pr(E)} \\
&= \frac{\Pr(E \cap F)}{\Pr(E)} \\
&= \frac{\Pr(E|F) \cdot \Pr(F)}{\Pr(E)} \\
&= \frac{\Pr(E) \cdot \Pr(F)}{\Pr(E)} \\
&= \Pr(F).
\end{aligned}$$

Intuitively, this means that if knowledge of $F$ doesn't tell you anything about $E$, then likewise knowledge of $E$ doesn't tell you anything about $F$.

**Example 6.12.**
Suppose two distinguishable, fair, six-sided dice are rolled; say one die is red and the other is blue. For each $i$ and $j$ between 1 and 6, let $B_i$ denote the event we roll $i$ on the blue die regardless of what happens with the red die; and $R_j$ is the event we roll $j$ on the red die, regardless of what the blue die rolls. (E.g., writing the result of the dice roll as a pair $(i, j)$ where $i$ is the value of the die and $j$ the value of the red die, the event $B_3$ is the event $\{(3,1),(3,2),(3,3),(3,4),(3,5),(3,6)\}$, and $R_1$ is the event $\{(1,1),(2,1),...,(6,1)\}$.) Are $B_i$ and $R_j$ independent events?

Intuitively we would expect these events to be independent since one does not influence the other, but let's check this is the case using our definition of conditional probability. Notice no matter what $i$ and $j$ are, $\Pr(B_i)$ and $\Pr(R_j)$ are both $6/36 = 1/6$ since there are 36 possible outcomes of rolling the two dice, and six of them correspond to rolling $i$ on blue or $j$ on red. Similarly, for all values of $i$ and $j$ we have $\Pr(B_i \cap R_j) = 1/36$ since there's only one way to get $i$ on blue *and* $j$ on red.

Now, for all $i$ and $j$ we have

$$\Pr(B_i|R_j) = {}^{\Pr(B_i \cap R_j)}/_{\Pr(R_j)} = \frac{{}^1/_{36}}{{}^1/_6} = {}^6/_{36} = {}^1/_6.$$

That is, $\Pr(B_i|R_j) = \Pr(B_i)$ and so the events are independent.

It's always good to have some intuition when trying to solve problems, but you have to be careful because sometimes your intuition can lead you astray. The next example shows us that two events can still satisfy our definition of independence, even when the events do seem like they should influence one another.

**Example 6.13.**
Let $F$ be the event we roll four or less on a fair, six-sided die, and let $E$ be the event we roll an even number. Are $E$ and $F$ independent?

Note first that $\Pr(F) = {}^4/_6 = {}^2/_3$, $\Pr(E) = {}^1/_2$, and $\Pr(E \cap F) = {}^2/_6 = {}^1/_3$. (For the last probability, notice $E \cap F$ is the event we roll an even number and we roll four or less. The only options are rolling 2 or 4, so two of the six possible outcomes.) Now applying the definition of conditional probability,

$$\Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{{}^1/_3}{{}^2/_3} = {}^1/_2$$

But $\Pr(E) = {}^1/_2$, so $\Pr(E|F) = \Pr(E)$ and the events *are* independent.

**Exercise 6.6.**
Suppose $E$ and $F$ are two events which each have non-zero probability ($\Pr(E) > 0$ and $\Pr(F) > 0$), but which are disjoint $E \cap F = \emptyset$. Are $E$ and $F$ independent?

It is sometimes convenient in mathematics to know that one concept can be described in several equivalent ways. For example, you know from

calculus that saying is a differentiable function is increasing is equivalent to saying the derivative of the function is positive. This means that if you're interested in seeing where a function is increasing, you have two options: you can either try to find intervals where for every $x_1$ and $x_2$ in the interval which satisfy $f(x_2) \geq f(x_1)$ whenever $x_2 \geq x_1$, or you can find the values of $x$ that solve $f'(x) \geq 0$. I.e., you can use the definition of increasing, or you can use some equivalent fact. Being aware of these equivalent facts gives you more tools you can use to solve problems, and this is nice because some tools might be easier to use than others.

For independents events in a same space, we have the following condition which is equivalent to our original definition of independence:

**Lemma 6.2.**
*Two events $E$ and $F$ in a sample space $\Omega$ are independent if and only if*

$$\Pr(E \cap F) = \Pr(E) \cdot \Pr(F).$$

*Proof.*
Since the statement here is *if and only if*, there are actually two things we have to check. We must first show that if $E$ and $F$ are independent, then it follows that $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$; and we must also show that if $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$, then $E$ and $F$ must be independent.

Suppose first that $E$ and $F$ are independent. By definition, this means $\Pr(E|F) = \Pr(E)$. Now, even if $E$ and $F$ weren't independent, we know that $\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F)$ by doing some algebra to the definition of conditional probability. Hence if $E$ and $F$ *are* independent, then

$$\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F) = \Pr(E) \cdot \Pr(F)$$

because $\Pr(E|F) = \Pr(E)$.

For the other direction, suppose we know $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$ and we want to show $\Pr(E|F) = \Pr(E)$. We again use the fact $\Pr(E \cap F) = \Pr(E|F) \cdot \Pr(F)$ (this *always* works whether $E$ and $F$ are

independent or not). We then have the following string of implications:

$$\begin{aligned} & \Pr(E \cap F) = \Pr(E) \cdot \Pr(F) \\ \implies & \Pr(E|F) \cdot \Pr(F) = \Pr(E) \cdot \Pr(F) \\ \implies & \frac{\Pr(E|F) \cdot \Pr(F)}{\Pr(F)} = \frac{\Pr(E) \cdot \Pr(F)}{\Pr(F)} \\ \implies & \Pr(E|F) = \Pr(E). \end{aligned}$$

□

**Exercise 6.7.**
Can an event be independent of itself? If this is possible, what can be said about the events which are independent of themselves?

The lemma above tells us that checking if $\Pr(E \cap F)$ equals $\Pr(E) \cdot \Pr(F)$ or not is just as good as checking if $E$ and $F$ are independent events. Thus, we can actually take $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$ as our definition of independence. This is useful if we want to extend the definition of independence to more than two events.

We say a finite collection of events $E_1$, $E_2$, ..., $E_n$ are ***mutually independent*** if for every collection of $1 \leq k \leq n$ subsets of $\{E_1, E_2, ..., E_n\}$ – say $E_{i_1}$, $E_{i_2}$, ..., $E_{i_k}$ – we have

$$\Pr(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = \Pr(E_{i_1}) \cdot \Pr(E_{i_2}) \cdot \ldots \cdot \Pr(E_{i_k}).$$

This definition looks a little strange the first time you see it, so let's write down explicitly what this means for two events, three events, and four events.

In the case of $n = 2$, so we are only considering a collection of two events $E_1$ and $E_2$, this definition simply says the events are mutually independent if

$$\Pr(E_1 \cap E_2) = \Pr(E_1) \cdot \Pr(E_2).$$

That is, for two events this is really just our normal notion of independence.

If we have three events, $E_1$, $E_2$, and $E_3$, then the definition says the following four conditions have to be satisfied for those three events to be

mutually independent:

$$\begin{aligned}
\Pr(E_1 \cap E_2) &= \Pr(E_1) \cdot \Pr(E_2) \\
\Pr(E_1 \cap E_3) &= \Pr(E_1) \cap \Pr(E_3) \\
\Pr(E_2 \cap E_3) &= \Pr(E_2) \cap \Pr(E_3) \\
\Pr(E_1 \cap E_2 \cap E_3) &= \Pr(E_1) \cdot \Pr(E_2) \cdot \Pr(E_3).
\end{aligned}$$

So, for three events to be mutually independent we need that each pair of events is independent (this is the first three equations above), but additionally the probability of the intersection of all of the events must equal the product of the probabilities of the three individual events.

In the case of four events $E_1$, $E_2$, $E_3$, and $E_4$, we have the following eleven conditions:

$$\begin{aligned}
\Pr(E_1 \cap E_2) &= \Pr(E_1) \cdot \Pr(E_2) \\
\Pr(E_1 \cap E_3) &= \Pr(E_1) \cdot \Pr(E_3) \\
\Pr(E_1 \cap E_4) &= \Pr(E_1) \cdot \Pr(E_4) \\
\Pr(E_2 \cap E_3) &= \Pr(E_2) \cdot \Pr(E_3) \\
\Pr(E_2 \cap E_4) &= \Pr(E_2) \cdot \Pr(E_4) \\
\Pr(E_3 \cap E_3) &= \Pr(E_3) \cdot \Pr(E_4) \\
\Pr(E_1 \cap E_2 \cap E_3) &= \Pr(E_1) \cdot \Pr(E_2) \cdot \Pr(E_3) \\
\Pr(E_1 \cap E_2 \cap E_4) &= \Pr(E_1) \cdot \Pr(E_2) \cdot \Pr(E_4) \\
\Pr(E_1 \cap E_3 \cap E_4) &= \Pr(E_1) \cdot \Pr(E_3) \cdot \Pr(E_4) \\
\Pr(E_2 \cap E_3 \cap E_4) &= \Pr(E_2) \cdot \Pr(E_3) \cdot \Pr(E_4) \\
\Pr(E_1 \cap E_2 \cap E_3 \cap E_4) &= \Pr(E_1) \cdot \Pr(E_2) \cdot \Pr(E_3) \cdot \Pr(E_4)
\end{aligned}$$

A succinct way to say this is that for four events to be mutually independent, we need that all collections of three events are mutually independent (this is the first ten conditions), and then there's one more condition that the probability of the quadruple intersection is the product of the probabilities of all of the events.

**Example 6.14.**
Imagine we have two distinguishable fair, six-sided dice, say one is blue and one is red. We roll both dice simultaneously and record the values on each die. Let $B_3$ be the event that we roll three on the blue die; $R_4$ the event we roll four on the red die; and $S_7$ the event that the

sum of the two dice is seven.

(a) Are all pairs of two events ($B_3$ and $R_4$; $B_3$ and $S_7$; $R_4$ and $S_7$) independent?

(b) Are all three events mutually independent?

---

(a) Let's first note that $\Pr(B_3) = 1/6$ and $\Pr(R_4) = 1/6$, and making a table of all the ways to roll two dice and get a sum of seven will show $\Pr(S_7) = 1/6$ as well:

| **Blue** | **Red** |
|---|---|
| 6 | 1 |
| 5 | 2 |
| 4 | 3 |
| 3 | 4 |
| 2 | 5 |
| 1 | 6 |

So, of the thirty-six ways to roll the two dice, six correspond to having a sum of seven, and $\Pr(S_7) = 6/36 = 1/6$. Now let's compute conditional probabilities to see if the events are independent or not:

$$\Pr(B_3|R_4) = \frac{\Pr(B_3 \cap R_4)}{\Pr(R_4)} = \frac{1/36}{1/6} = \frac{1}{6} = \Pr(B_3)$$
$$\Pr(B_3|S_7) = \frac{\Pr(B_3 \cap S_7)}{\Pr(S_7)} = \frac{1/36}{1/6} = \frac{1}{6} = \Pr(B_3)$$
$$\Pr(R_4|S_7) = \frac{\Pr(R_4 \cap S_7)}{\Pr(S_7)} = \frac{1/36}{1/6} = \frac{1}{6} = \Pr(R_4)$$

Thus all pairs of the three events are independent: each of our three events is independent for each other one.

(b) Mutual independence requires the three conditions from part (a) to be satisfied, and one more condition. We need to see if $\Pr(B_3 \cap R_4 \cap S_7)$ equals $\Pr(B_3) \cdot \Pr(R_4) \cdot \Pr(S_7)$ or not. Notice

$$\Pr(B_3 \cap R_4 \cap S_7) = \frac{1}{36},$$

however

$$\Pr(B_3) \cdot \Pr(R_4) \cdot \Pr(S_7) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{216}$$

Hence the three events *are not* mutually independent!

## 6.8 Relating probabilities of intersections and unions

At this point we have some formulas for computing probabilities of unions and intersections of events, provided the events satisfy some special conditions.

For example, we know that if a collection of events $E_1$, $E_2$, ..., $E_n$ are pairwise disjoint (i.e., if $E_i \cap E_j = \emptyset$ for any two distinct indices $i$ and $j$), then

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_n) = \Pr(E_1) + \Pr(E_2) + \cdots + \Pr(E_n).$$

We also know that if a collection of events $E_1$, $E_2$, ..., $E_n$ are mutually independent, then

$$\Pr(E_1 \cap E_2 \cap \cdots \cap E_n) = \Pr(E_1) \cdot \Pr(E_2) \cdot \ldots \cdot \Pr(E_n).$$

It can sometimes be convenient to relate these two different formulas, and de Morgan's laws (see Section 2.6) give us a way to relate unions and intersections by taking complements.

First we note that we can use de Morgan's laws to show that if $E$ and $F$ are independent events, then $E^c$ and $F^c$ must be independent as well.

**Lemma 6.3.**
*If $E$ and $F$ are two independent events in a sample space $\Omega$, then their complements $E^c$ and $F^c$ are independent as well.*

*Proof.*
Assume $E$ and $F$ are independent. We want to show $E^c$ and $F^c$ are independent by checking if $\Pr(E^c|F^c)$ equals $\Pr(E^c)$. We will simply write out the definition of conditional probability, use de Morgan's laws to turn the intersection into a union, and then apply inclusion-exclusion to find the probability of that union, and finally just do some basic algebra to manipulate this expression.

$$\begin{aligned}
\Pr(E^c|F^c) &= \frac{\Pr(E^c \cap F^c)}{\Pr(F^c)} \\
&= \frac{\Pr((E \cup F)^c)}{\Pr(F^c)} \\
&= \frac{1 - \Pr(E \cup F)}{\Pr(F^c)} \\
&= \frac{1 - (\Pr(E) + \Pr(F) - \Pr(E \cap F))}{\Pr(F^c)} \\
&= \frac{1 - (\Pr(E) + \Pr(F) - \Pr(E) \cdot \Pr(F))}{\Pr(F^c)} \\
&= \frac{1 - (\Pr(F) + \Pr(E) \cdot (1 - \Pr(F)))}{\Pr(F^c)} \\
&= \frac{1 - \Pr(F) - \Pr(E) \cdot (1 - \Pr(F))}{\Pr(F^c)} \\
&= \frac{\Pr(F^c) - \Pr(E)\Pr(F^c)}{\Pr(F^c)} \\
&= \frac{\Pr(F^c) \cdot (1 - \Pr(E))}{\Pr(F^c)} \\
&= 1 - \Pr(E) \\
&= \Pr(E^c)
\end{aligned}$$

□

The previous lemma extends to mutually independent events and the proof is simply induction with the base case being independence of two events, which is supplied by Lemma 6.3. We won't give the proof here because it's tedious to write down, but that's the basic idea.

**Lemma 6.4.**
*If $E_1$, $E_2$, ..., $E_n$ are mutually independent events in a sample space $\Omega$, then the complements $E_1^c$, $E_2^c$, ..., $E_n^c$ are also mutually independent.*

This is convenient because de Morgan's laws tell us that unions become intersections when we take complements, and if the events are mutually independent then we can calculate the probability of the intersection easily.

**Corollary 6.5.**
*If $E_1$, $E_2$, ..., $E_n$ are mutually independent events in a sample space $\Omega$, then*

$$\Pr(E_1 \cup E_2 \cup \cdots \cup E_n) = 1 - \Pr(E_1)^c \cdot \Pr(E_2)^c \cdot \ldots \cdot \Pr(E_n)^c$$

**Exercise 6.8.**
Prove Corollary 6.5 by writing $\Pr(E_1 \cup E_2 \cup \cdots \cup E_n)$ in terms of its complement, applying de Morgan's law, and then using Lemma 6.4.

**Example 6.15.**
Suppose a seam on the wing of a certain type of aircraft has twenty-five rivets. For safety reasons, if any one of the rivets is defective then the whole seam has to be reworked. Assume the event a rivet is defective is mutually independent from all the other rivets being defective. If 20% of seams have to be reworked, what is the probability an individual rivet is defective?

Let $R_i$ denote the event the $i$-th rivet ($1 \leq i \leq 25$) is defective. We're told in the statement of the problem that the probability any one rivet on the seam is defective (i.e., if rivet 1 is defective, or rivet

2 is defective, or rivet 3 is defective, ...) is 0.2:

$$\Pr(R_1 \cup R_2 \cup \cdots \cup R_{25}) = 0.2.$$

From this we want to compute $\Pr(R_i)$. Since the rivets are mutually independent, we can apply Corollary 6.5 to rewrite $\Pr(R_1 \cup \cdots \cup R_{25})$ as

$$\begin{aligned}\Pr(R_1 \cup \cdots \cup R_{25}) &= 1 - \Pr([R_1 \cup \cdots \cup R_{25}]^c) \\ &= 1 - \Pr(R_1^c \cap \cdots \cap R_{25}^c) \\ &= 1 - \Pr(R_1)^c \cdot \ldots \cdot \Pr(R_{25}^c)\end{aligned}$$

Now, assuming all rivets have the same probability of being defective, we can write this as

$$\Pr(R_1 \cup \cdots \cup R_{25}) = 1 - (\Pr(R_i)^c)^{25}$$

for any rivet. We know this equals 0.2, however, and so

$$\begin{aligned}1 - (\Pr(R_i^c))^{25} &= 0.2 \\ \implies -(\Pr(R_i^c))^{25} &= -0.8 \\ \implies (\Pr(R_i^c))^{25} &= 0.8 \\ \implies \Pr(R_i^c) &= \sqrt[25]{0.8} \\ \implies \Pr(R_i) &= 1 - \sqrt[25]{0.8} \approx 0.008886\end{aligned}$$

So, an individual rivet is defective about 0.89% of the time.

**Remark.**
Notice that the probability of a defective rivet in the example above *is not* 0.008, which would be $0.2/25$. That is, $\Pr(R_1 \cup \cdots \cup R_{25}) \neq 25 \Pr(R_i)$. The issue of course is that these events are not mutually disjoint: rivet 1 and rivet 2 can both be defective, meaning $R_1 \cap R_2 \neq \emptyset$. In fact, if two events are independent (such as our $R_i$ events in the example above) and have positive probability, they *can not* be disjoint by Exercise 6.6.

## 6.9 Practice problems

**Problem 6.1.**
Suppose that one morning while getting ready for class you are in a hurry, not paying attention to what you're doing, and simply reach into your sock drawer and pull out two random socks. Supposing your sock drawer has twelve white socks, six black socks, four brown socks, and four blue socks. (These are individual socks, not pairs.)

What is the probability both socks are blue, given that you grabbed two socks of the same color?

**Problem 6.2.**
Suppose that $E$ and $F$ are two events in some sample space $\mathcal{S}$ where $P(E) = 2/5$, $P(F) = 3/10$, and $P(E \cup F) = 1/2$. What is $P(E|F)$?

**Problem 6.3.**
Suppose an urn contains 100 marbles which are labelled 1 through 100. You reach into the urn and pull out one marble. Let $E$ be the event that the marble you pulled out has a label which is an even number, and $F$ the event the marble you pull out has a label which is a multiple of five. Are $E$ and $F$ independent events?

**Problem 6.4.**
At a certain high school, 30% of students play soccer, 10% play football, and 25% play basketball. Suppose that, of the students that play both football and soccer, 5% also play basketball; of the students that play soccer, 10% play football.

What is the probability a randomly selected student plays all three sports?

**Problem 6.5.**
Suppose that the population of the United Kingdom is split up as follows: 60 percent of the population is English, 20 percent is Scottish, 15 percent is Northern Irish, and 5 percent is Welsh. Of these, 15% of the English have red hair, 75% of the Scottish have red hair, 65% of the Northern Irish red hair, and 30% of the Welsh have red hair.

If a random redhead from the UK is selected, what is their nationality most likely to be? What is the second most likely nationality?

**Problem 6.6.**
Suppose a particular course is a requirement for all math majors, but can not be taken by freshmen. Suppose that 20% of math majors are freshmen, 30% are sophomores, 25% are juniors, and 25% are seniors. Suppose also

that one third of sophomores are currently enrolled in the course, half of juniors are currently enrolled in the course, and one fourth of seniors are enrolled in the course.

(a) What percentage of math majors are currently enrolled in the course?

(b) What is the probability a randomly selected student in the course is a junior?

**Problem 6.7.**
Suppose 15% of mathematics majors at a given university go on to work in finance; 5% of computer science majors at this university go into finance; 10% of physics majors go into finance; and 3% of students from other majors ultimately go into finance. Assume that 5% of the students at this university major in mathematics, 10% major in computer science, and 3% major in physics. If a random graduate of this university working in finance is selected, what is the probability they were a mathematics major?

**Problem 6.8.**
A manufacturer of running shoes has plants in South Korea, Australia, and Venezuela. South Korea produces 60 percent of the shoes, Australia 20 percent, and Venezuela 20 percent. They make 2 types of shoes at each plant, a racing shoe and a training flat. The production at each plant is allocated as shown in the table. Suppose that these shoes are randomly distributed in stores in the United States and that you go into a store and buy a training flat. What is the probability that is came from South Korea?

| **Plant** | **Racing** | **Training** |
|---|---|---|
| South Korea | 0.5 | 0.5 |
| Australia | 0.25 | 0.75 |
| Venezuela | 0.4 | 0.6 |

# Part III

# Random Variables

# 7

# *Introduction to Random Variables*

*La mathématique est l'art de donner le même nom à des choses différentes.*
Mathematics is the art of giving the same name to different things.

---

Henri Poincaré
*L'avenir des mathématiques*

## 7.1 The idea of a random variable

In many experiments we don't really care about the exact outcome of the experiment, but rather some quantitative value determined by that outcome. For example, if we imagine playing a game of darts on a normal dart board which is divided up into regions worth various points, we don't really care what exact point our dart hits; what we care about is the number of points we get. In a game where you roll dice, you may not care about the exact values of the dice you roll, but only the sum of those values. (E.g., if you roll three dice you may care that you roll a total of 13, but whether that's from rolling $(6, 3, 4)$, or $(5, 2, 6)$, or $(1, 6, 6)$ doesn't really matter.) The mathematical formulation of this idea of having a numerical value determined by a random experiment is called a random variable.

To be precise, a ***random variable*** associated to an experiment with sample space $\Omega$ is simply a function whose domain (the set of inputs to the function) is $\Omega$ and whose codomain (the set of possible outputs) is the set of real numbers, $\mathbb{R}$. We usually denote random variables with capital letters towards the end of the alphabet like $X$, $Y$, or $Z$.

The name *random variable* may seem strange as there is nothing "random" about the function itself – it's a normal, deterministic function. The idea, though, is that we will perform our experiment and plug the outcome of that experiment (which is random) into the function. In this way we have a random number that depends on the outcome of our experiment.

**Example 7.1.**
Suppose three coins are flipped simultaneously. The sample space of

this experiment consists of eight simple events,

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

Consider the function $X : \Omega \to \mathbb{R}$ which associates to each simple event the number of heads:

$$\begin{aligned} X(\texttt{TTT}) &= 0 \\ X(\texttt{TTH}) &= 1 \\ X(\texttt{THT}) &= 1 \\ X(\texttt{THH}) &= 2 \\ X(\texttt{HTT}) &= 1 \\ X(\texttt{HTH}) &= 2 \\ X(\texttt{HHT}) &= 2 \\ X(\texttt{HHH}) &= 3 \end{aligned}$$

Thus we have a function which takes an outcome of our experiment and associates to that outcome a number. Since we don't know what the outcome of the experiment will be beforehand, we don't know what the output of our function will be. In this way the function gives us a random number.

**Example 7.2.**
Suppose we take one coin and flip it repeatedly until we get a heads on the coin, and consider the random variable that tells us the number of flips required to get that first heads. The sample space of this experiment is

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

Our random variable $X : \Omega \to \mathbb{R}$ in this case would be

$$\begin{aligned} X(\texttt{H}) &= 1 \\ X(\texttt{TH}) &= 2 \\ X(\texttt{TTH}) &= 3 \\ X(\texttt{TTTH}) &= 4 \\ X(\texttt{TTTTH}) &= 5 \\ &\vdots \end{aligned}$$

**Example 7.3.**
The acidity of soil can affect the health of any vegetables grown in that soil, and for this reason a farmer may be interested in measuring how acidic the soil in their field is. To test this, the farmer may go into their field, take a sample sample of soil, and then determine the pH of the soil using a pH meter. We can interpret this as a random variable: the experiment being performed is taking a random sample of soil from the field, and the number we associate to the outcome of this experiment is the measured pH. (By definition, pH is a number between 0 and 14.)

**Example 7.4.**
Imagine a factory produces lightbulbs which fail after a certain lifetime. (That is, the lightbulbs don't last forever. They burn for a while, and then go out. The amount of time the lightbulb lasts, meaning the number of hours the lightbulb is turned on before it burns out, is the lifetime.) We may perform an experiment where we select a random lightbulb produced by this factory, turn it on, and then record how long it the lightbulb lasts before burning out. The sample space of this experiment would be the set of all lightbulbs produced by the factory, and the random variable is the function which associates to each lightbulb its lifetime.

## 7.2 Discrete versus continuous

For our purposes, we will think about random variables being separated into two basic types called *discrete* and *continuous*. The difference between the two depends on the range of the random variable.

**Remark.**
The codomain of our random variables is always the set of all real numbers, but the range is the set of real numbers that area actually acheived as outputs of the function.

A random variable $X : \Omega \to \mathbb{R}$ is called ***discrete*** if either of the following conditions is satisfied:

- The range of $X$ is finite, or
- The range of $X$ is infinite, but the values in the range can be ordered as the first value, second value, third value, and so on.

The random variables in Example 7.1 and Example 7.2 are both discrete. In Example 7.1 the range is finite, whereas in Example 7.2 the range is infinite but there's a well-defined first outcome, second outcome, third outcome, and so on.

A random variable $X : \Omega \to \mathbb{R}$ is called ***continuous*** if the range of the function is an interval of the real line. Here we allow the interval to be closed (such as $[a, b]$), or open (such as $(a, b)$), or half-open (such as $[a, b)$); it can be bounded (such as $[a, b]$), or unbounded (such as $(a, \infty)$, $(-\infty, b]$, $(-\infty, \infty)$, ...). The important thing is that there is an infinite range of values, and this range is too big to say there's a first element, a second element, and so on.

In Example 7.3, the pH measured can be anything in the interval $[0, 14]$, and so this is a continuous random variable. In Example 7.4, the lifetime of a lightbulb is a continuous random variable with values in the interval $[0, \infty)$ if we believe the lifetime of a lightbulb could be arbitrarily long. If we had some reason to know that a lightbulb can't last more than, say, 10 years, then we could limit the range of values to $[0, 87600]$. Either way, this is also a continuous random variable. We

Ultimately we are interested in computing the probability a random variable will take on a given value or a range of values. E.g., we may want

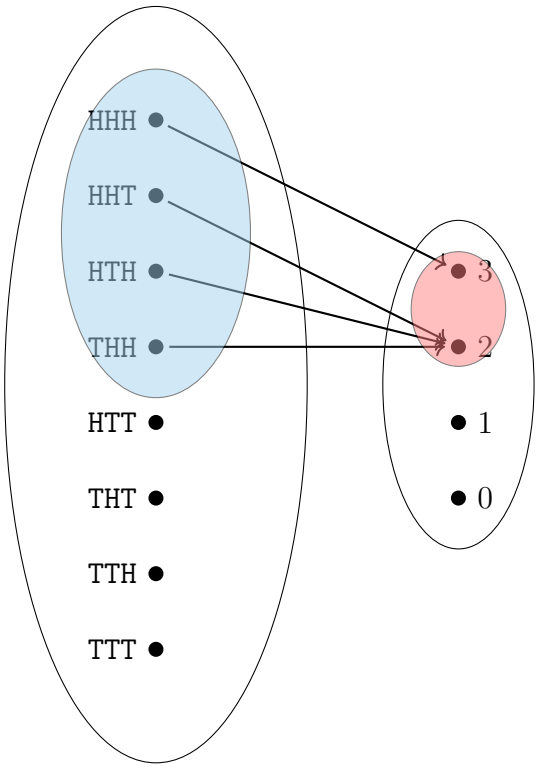
to know the probability we flip at least two heads, or the probability the pH of a soil sample is between 5.8 and 6.8. We will first discuss how to compute these probabilities in principle, in a way that will work for both discrete and continuous random variables. Later on we will see there are often simpler methods for specific types of random variables, but let's get the basic idea out of the way first.

Recall from Section 3.4 that given any function $f : A \to B$ between two sets, we can associate a subset of $A$ to each subset of $B$. In particular, if $E \subseteq B$, the ***preimage*** of $E$ under the function $f : A \to B$ is the set of all elements in $A$ which $f$ maps to an element of $E$. This set is (somewhat unfortunately) denoted $f^{-1}(E)$:

$$f^{-1}(E) = \{a \in A \,\big|\, f(a) \in E\}.$$

To have a concrete example, consider the random variable from Example 7.1. To each outcome, we associate a number by counting the number of heads. We can think of this as a function $X : \Omega \to \mathbb{R}$. The preimage of the set $\{2, 3\} \subseteq \mathbb{R}$, denoted $X^{-1}(\{2, 3\})$, is the set of outcomes which result in two or three heads.

In the figure below, the red region represents the set $\{2, 3\}$, and the blue region represents its preimage.



**Remark.**
Notice that the preimage could be empty! In the example above, $X^{-1}(\{4\})$ would be the empty set since there's no way to flip three coins and get four heads.

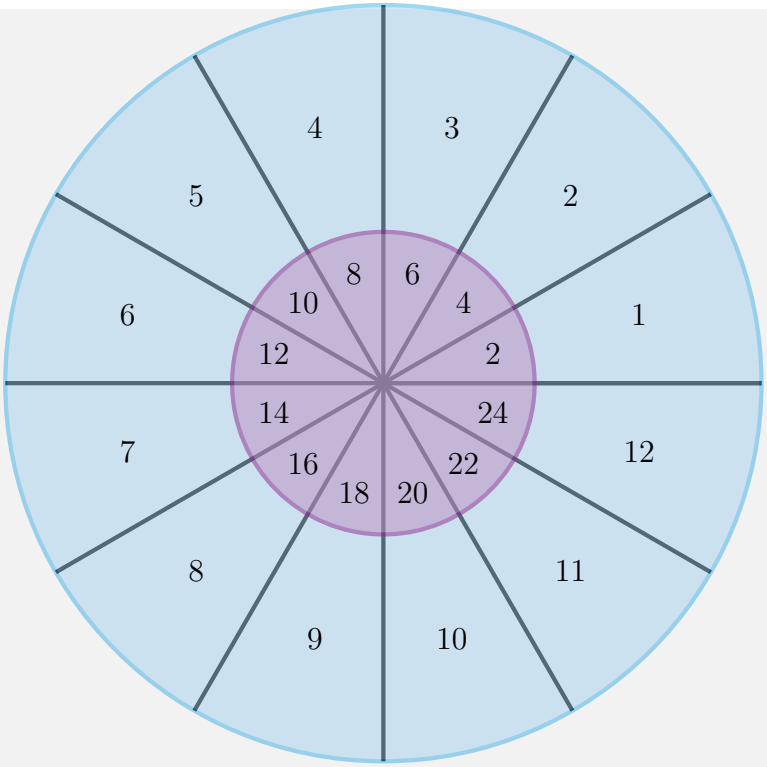
Now notice that $X^{-1}(\{2,3\})$ is a subset of the sample space $\Omega$, and so it's something we can calculate the probability of. The probability the output of $X$ is 2 or 3 is $\Pr(X^{-1}(\{2,3\})$ which we know is $1/2$.

## 7.3 Probability the random variable takes on a given value

Given a random variable $X$ associated to an experiment with sample space $\Omega$, the probability $X$ gives us a value in the set $E \subseteq \mathbb{R}$ is $\Pr(X^{-1}(E))$. Thus we convert sets of real numbers we are interested in into events in the sample space whose probabilities we can compute.

Sometimes we will be a little bit lazy with the notation and write $\Pr(X^{-1}(E))$ simply as $\Pr(X \in E)$, or $\Pr(X = k)$ for $\Pr(X^{-1}(\{k\}))$, or $\Pr(a \leq X \leq b)$ for $\Pr(X^{-1}([a,b]))$, but the idea is always the same: to find the probability our random variables gives us a number in a given subset of the real line, we look at the preimage of that set to get an event in our sample space and then calculate the probability of that event. That is, to find the probability the random variable $X$ gives us a value in the interval $[a,b]$ (i.e., the probability $a \leq X \leq b$) we look at the set of all outcomes of the experiment which $X$ sends into the interval (i.e., the preimage $X^{-1}([a,b])$), and compute the probability of this event. Technically should be denoted $\Pr(X^{-1}([a,b]))$, but sometimes it's convenient to write this as $\Pr(a \leq X \leq b)$ since this really describes what we're interested in: the probability the random variable $X$ takes on a value between $a$ and $b$.

**Example 7.5.**
Consider throwing darts randomly at a dart board where the board is separated into regions and if the dart lands in a given region, the player is awarded a specified number of points. Suppose the dart board is broken into these regions as indicated below.




The player scores four points, for example, if they land in either of the two sectors labelled by 4 (one is a large blue region and one is a small purple region). How do we find the probability the player scores a particular value? Assuming the player is just as likely to hit any point on the dartboard as any other point, we find the area of the region which gives the the given number of points, divided by the total area of the board.

To be concrete, let's suppose the dart board has a radius of one unit, and the inner purple circle has a radius of $\sqrt{1/3}$ units. Since the board is divded into twelve sectors, the area of each sector (which consists of a blue part and a purple part) is $\pi/12$. Thus the area of each purple region is $\pi/36$, and the area of each blue region is $\pi/18$. (Since the purple and blue regions are $1/3$ and $2/3$ of each sector, respectively.)

The probability the player scores four points is then equal to

$$\frac{\pi/36 + \pi/18}{\pi} = \frac{3}{36} = \frac{1}{12},$$

and the probability the player scores three points is

$$\frac{\pi/18}{\pi} = \frac{1}{18},$$

and the probability the player scores twenty points is

$$\frac{\pi/36}{\pi} = \frac{1}{36}.$$

Here the random variable is the function that associates to each point on the board the number of points the player scores when their dart lands at that point. To find the probability of getting a certain score, we look at the the set of all points on the board which give us that score (these are the blue and purple wedges above). These wedges are some special events in our sample space, and so we can compute the probability of these events, and this gives us the probability of getting a particular score.

**Exercise 7.1.**
Is the random variable indicated in Example 7.5 discrete or continuous?

## 7.4 Practice problems

**Problem 7.1.**
Consider the following experiment: we roll a ball down a road and measure how far the ball travels. This measurement of the distance the ball travels is a random variable. Is this random variable discrete or continuous?

**Problem 7.2.**
Suppose we roll a ball down a road and measure how far the ball travels, rounded to the nearest integer. For instance, if the ball rolls 15.6723 feet, we round that up to 16. Is this random variable discrete or continuous?

**Problem 7.3.**
Imagine a game of darts where you score points based on how close the dart lands to the bullseye of the dart board. If the dart lands at the bullseye you are awarded 100 points, and the points scored decrease linearly down to 0 points if the dart hits the edge of the board. (If the board has radius $r$ and your dart lands distance $d$ from the origin, this means the number of points scored is $100\frac{r-d}{r}$.) Is this random variable discrete or continuous?

# 8

# *Discrete Random Variables*

*There should be no such thing as boring mathematics.*

EDSGER DIJKSTRA

In the last chapter we mentioned that, for our purposes in this class, we think of random variables as coming in two different flavors: discrete and continuous. Though the basic ideas of random variables are the same regardless of which type of random variable we have, in practice the way we do computations with these random variables often depends on whether the random variable is discrete or continuous. In this chapter we focus on discrete random variables and will turn our attention to continuous random variables later.

## 8.1 The probability mass function

Suppose $X : \Omega \to \mathbb{R}$ is a discrete random variable. If we only care about the value of the random variable and not the underlying experiment, we can think of the random variable as giving us a way of choosing a random real number, and we might like to know the probability of choosing a particular real number. We can introduce a function called the *probability mass function* of $X$ which tells us exactly this information. To be precise, the ***probability mass function***, often abbreviated ***pmf***, of a discrete random variable $X : \Omega \to \mathbb{R}$ is a function $p : \mathbb{R} \to \mathbb{R}$ defined as follows:

$$p(x) = \Pr(X = x) = \Pr(X^{-1}(x)).$$

That is, $p(x)$ tells us the probability the random variable $X$ will output a given real number $x$.

**Example 8.1.**
Consider the random variable which counts the number of heads that are seen when a fair coin is flipped three times, as in Example 7.1 from the previous chapter. We have already seen that if $\Omega$ is the set of all possible three-flip sequences of the coin, then $X : \Omega \to \mathbb{R}$ is the

function

$$\begin{aligned} X(\texttt{TTT}) &= 0 \\ X(\texttt{TTH}) &= 1 \\ X(\texttt{THT}) &= 1 \\ X(\texttt{THH}) &= 2 \\ X(\texttt{HTT}) &= 1 \\ X(\texttt{HTH}) &= 2 \\ X(\texttt{HHT}) &= 2 \\ X(\texttt{HHH}) &= 3 \end{aligned}$$

The probability mass function is the function which tells us how likely each possible output is. Since the sample space consists of eight equally-likely simple events, we can determine the pmf by simply counting the number of times each output appears above. That is, $p(0) = 1/8$ since only one of the eight possible simple events has zero heads; $p(1) = 3/8$ since there are three ways to get one head; $p(2) = 3/8$ since there are three ways to get two heads; and $p(3) = 1/8$ since there's only one possible way to get three heads.

How many ways are there to get four heads? Since there are only three coins being flipped, we can't get four heads so the probability of getting four heads is zero, and hence $p(4) = 0$. Similarly, $p(5) = 0$, $p(6) = 0$, $p(-2) = 0$, $p(\pi) = 0$, and so on.

The probability mass function of the random variable is thus the following function:

$$p(x) = \begin{cases} 1/8 & \text{if } x = 0 \\ 3/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 2 \\ 1/8 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

**Example 8.2.**
Consider an experiment where a fair coin is tossed repeatedly until it first lands on heads. Let $X$ be the random variable which counts the

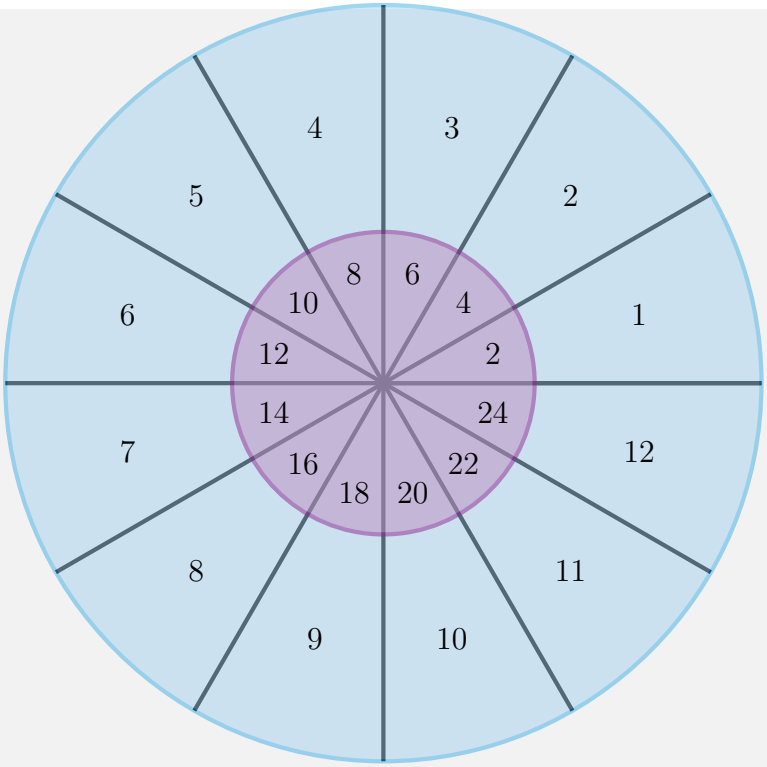
number of required flips until getting heads. What is the pmf of $X$?

For each natural number $n \in \mathbb{N}$, the pmf is defined by $p(n) = \Pr(X^{-1}(n))$. I.e., $p(n)$ is the probability of getting the first heads on the $n$-th flip. Since there are $2^n$ ways a coin can be flipped $n$ times (two possible outcomes for each of the $n$ flips), and only one one of these corresponds to getting the first heads. Thus the probability we get the first heads on the $n$-th flip is $1/2^n$. Thus

$$p(x) = \begin{cases} 1/2^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

**Exercise 8.1.**
Suppose a coin is weighted so that its probability of coming up heads is different from its probability of coming up tails; say the probability of heads is $2/3$ and the probability of heads is $1/3$. This coin is repeatedly flipped until the first heads comes up, and the random variable $X$ records the number of required flips until the first heads. What is the pmf of $X$? (Hint: Start off by thinking of small values of $X$. What is the probability $X = 1$? What about $X = 2$ and $X = 3$? Once you understand the small values, see if there's a pattern you can generalize.)

**Example 8.3.**
In Example 7.5 we considered throwing darts at a dart board which was divided into regions worth various points as indicated below.




To find the pmf of this random variable we consider the area of the regions which can score a given number of points. As discussed in Example 7.5, if the board has radius 1, then the area of each blue region is $\pi/18$ and the area of each purple region is $\pi/36$. Hence the probability of scoring a number that is represented by only a blue region (this consists only of the odd numbers 1, 3, 5, ..., 11) is $1/18$; the probability of scoring a number that is represented by a blue region and a purple region (these are the even numbers 2, 4, 6, ..., 12) is $1/12$; and the probability of scoring a number that is represented only by a purple region (these are the values 14, 16, 18, ..., 24) is $1/36$. Of course, we can not score points equal to a number not represented on the board, so the probability of any such number is zero. Putting all of this together, the pmf of the random variable which associates scores

to points on the board is

$$p(x) = \begin{cases} 1/18 & \text{if } x = 1 \\ 1/12 & \text{if } x = 2 \\ 1/18 & \text{if } x = 3 \\ 1/12 & \text{if } x = 4 \\ 1/18 & \text{if } x = 5 \\ 1/12 & \text{if } x = 6 \\ 1/18 & \text{if } x = 7 \\ 1/12 & \text{if } x = 8 \\ 1/18 & \text{if } x = 9 \\ 1/12 & \text{if } x = 10 \\ 1/18 & \text{if } x = 11 \\ 1/12 & \text{if } x = 12 \\ 1/36 & \text{if } x = 14 \\ 1/36 & \text{if } x = 16 \\ 1/36 & \text{if } x = 18 \\ 1/36 & \text{if } x = 20 \\ 1/36 & \text{if } x = 22 \\ 1/36 & \text{if } x = 24 \\ 0 & \text{otherwise} \end{cases}$$

Notice that if the pmf of a random variable is known, we can essentially forget about the underlying experiment: everything you need to know about the random variable is contained in its pmf. It is very common for us to work with random variables in this way, discussing only the pmf and completely ignoring any underlying experiment. For this reason it would be helpful to know some basic properties of the pmf that apply for all discrete random variables regardless of what that underlying experiment may be.

Before giving some basic properties, let's extend our definition of the pmf to subsets of $\mathbb{R}$ instead of just individual numbers. By definition, the pmf of a discrete random variable $X$ is given by the equation

$$p(x) = \Pr(X^{-1}(x)).$$

We can easily extend this to subsets of $\mathbb{R}$ by computing the probability of

the preimage of the subset. That is, for any $E \subseteq \mathbb{R}$ we define

$$p(E) = \Pr(X^{-1}(E)).$$

The way we should interpret this number is that it is the probability the random variable will give us some value in the set $E$. For example, using the pmf in the dart example above, $p([1,5])$ should be the probability that we score between 1 and 5 points. Adding up the areas of the blue and purple sectors which correspond to one, two, three, four, or five points, we see this is

$$p([1,5]) = 1/3.$$

The way we calculated this number was by looking at the probability of scoring one, two, three, four, or five points and adding these probabilities together – this is exactly the same as adding up the areas of the corresponding blue and purple sectors and dividing by the total area of the board.

The next theorem says, among other things, this type of calculation works for all discrete random variables and their pmf's.

**Theorem 8.1.**
*Suppose $X : \Omega \to \mathbb{R}$ is a discrete random variable with pmf $p(x)$. We then have the following properties:*

1. *For every $x \in \mathbb{R}$, $0 \leq p(x) \leq 1$.*
2. *For any infinite sequence of distinct real numbers $x_1$, $x_2$, $x_3$, ..., we have*
$$p\left(\bigcup_{n=1}^{\infty}\{x_n\}\right) = \sum_{n=1}^{\infty} p(x_n).$$
3. *$p(\mathbb{R}) = 1$*

*Proof.*

1. Recall that for any event $E$ in a sample space $\Omega$, $\Pr(E)$ is always between 0 and 1. Since for each real number $x$, $X^{-1}(x)$ is an event, we must have $\Pr(X^{-1}(x))$ is between 0 and 1. But

$\Pr(X^{-1}(x))$ is exactly $p(x)$ – this is the definition of $p(x)$ – and so $0 \leq p(x) \leq 1$.

2. Using our extended definition of $p(x)$,

$$p\left(\bigcup_{n=1}^{\infty}\{x_n\}\right) = \Pr\left(X^{-1}\left(\bigcup_{n=1}^{\infty}\{x_n\}\right)\right).$$

Since the elements of the sequence are assumed to be distinct, the events $X^{-1}(\{x_n\})$ are disjoint. Using the property that

$$X^{-1}\left(\bigcup_{n=1}^{\infty}\{x_n\}\right) = \bigcup_{n=1}^{\infty} X^{-1}(\{x_n\}),$$

we have a disjoint union of events in $\Omega$. One of the axioms of the probability function Pr is that the probability of a disjoint union of events equals the sum of the probabilities of the events, and so

$$\Pr\left(\bigcup_{n=1}^{\infty} X^{-1}(\{x_n\})\right) = \sum_{n=1}^{\infty} \Pr(X^{-1}(\{x_n\})).$$

The expression on the left equals $p\left(\bigcup_{n=1}^{\infty}\{x_n\}\right)$ while the expression othe right equals $\sum_{n=1}^{\infty} p(x_n)$, and so the result is proved.

3. By our extended definition of the pmf, $p(\mathbb{R})$ is $\Pr(X^{-1}(\mathbb{R}))$. Notice that $X^{-1}(\mathbb{R}) = \Omega$ since $X$ sends every element of $\Omega$ to some real number. Thus

$$p(\mathbb{R}) = \Pr(X^{-1}(\mathbb{R})) = \Pr(\Omega) = 1.$$

□

**Corollary 8.2.**

*For the pmf $p(x)$ of any discrete random variable $X$,*

$$\sum_{x\in\mathbb{R}} p(x) = 1.$$

*Proof.*
This is really just the third property of Theorem 8.1 but rewritten:

$$\sum_{x\in\mathbb{R}} p(x) = \sum_{x\in\mathbb{R}} \Pr(X^{-1}(x)) = \Pr\left(\bigcup_{x\in\mathbb{R}} \{x\}\right) = \Pr(\mathbb{R}) = 1.$$

□

**Exercise 8.2.**
Suppose $X$ is a random variable with the following pmf:

$$p(x) = \begin{cases} 1/15 & \text{if } x = -2 \\ 1/3 & \text{if } x = -1 \\ 2/15 & \text{if } x = 0 \\ 4/15 & \text{if } x = 1 \\ 1/5 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the probability $X$ is positive?

**Exercise 8.3.**
Verify the pmf from Exercise 8.1 satisfies $\sum_{x\in\mathbb{R}} p(x) = 1$.

(Hint: Recall there's a nice formula for the sum of a geometric series.)

## 8.2 The cumulative distribution function

The probability mass function of a discrete random variable tells us the probability the random variable takes on a given value. Sometimes we are instead interested in the probability the random variable takes on a range of values. To determine these probabilities we introduce a new function similar to the probability mass function called the *cumulative distribution function* of the random variable.

The ***cumulative distribution function***, or ***cdf***, of a discrete random variable $X$ is the function $F : \mathbb{R} \to \mathbb{R}$ which tells us the probability $X \leq x$ for each $x \in \mathbb{R}$. That is,

$$F(x) = \Pr(X \leq x) = \Pr(X^{-1}(-\infty, x]).$$

Again, this means that $F(x)$ tells us the probability $X$ is at most equal to $x$.

**Example 8.4.**
In the experiment where a coin is flipped three times and the number of heads is counted, we had seen in Example 8.1 that the pmf was

$$p(x) = \begin{cases} 1/8 & \text{if } x = 0 \\ 3/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 2 \\ 1/8 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

From this we can determine the cumulative distribution function, $F(x)$. Notice first that if $x < 0$, then $F(x) = 0$: the probability we have less than zero heads is zero. At $x = 0$, however, the value of $F(x)$ suddenly jumps to $1/8$ as $F(0)$ is the probability $X \leq 0$ and $X = 0$ with probability $1/8$.

For values of $x$ in the interval $(0, 1)$, $F(x)$ still equals $1/8$ since for each $x$ satisfying $0 \leq x \leq 1$, the probability $X \leq x$ includes the case $X = 0$ which has probability $1/8$.

Once $x$ reaches 1, value of $F(x)$ instantly jumps again to $1/2$. This is because if $X \leq 1$, that includes the both cases $X = 0$ and $X = 1$ and these occur with probabilities $1/8$ and $3/8$, respectively. Thus $\Pr(X \leq 1)$ equals $1/8 + 3/8 = 4/8 = 1/2$.

The value of $F(x)$ stays constant for $x$ beween 1 and 2, then at $x = 2$ suddenly jumps to $F(2) = 7/8$ since if $X \leq 2$, then $X$ could be 0, 1, or 2 and these possibilities occur with probabilities $1/8$, $3/8$, and $3/8$, and $1/8 + 3/8 + 3/8 = 7/8$.

The value of $F(x)$ again remains constant for $x$ between 2 and 3, then jumps at $x = 3$ to $F(3) = 1$. Note that for all $x \geq 3$ we have $F(3) = 1$ since if $x \geq 3$, $X \leq 3$ includes all possible values of $X$.

Putting all of the above together we have

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/8 & \text{if } 0 \leq x < 1 \\ 1/2 & \text{if } 1 \leq x < 2 \\ 3/8 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$

**Example 8.5.**
In the experiment of Example 8.2, a coin is flipped until heads appears and the random variable $X$ counts the required number of flips. We saw in that example that the pmf was given by $1/2^n$ for $X = n$, where $n$ was a positive integer, and zero everywhere else. This tells us that the cdf will be zero everywhere to the left of 1, and then instantly jumps to $1/2$ at $x = 1$. The cdf remains constains between 1 and 2, and then jumps to $1/2 + 1/4 = 3/4$ at $x = 2$. Similarly, the cdf remains constant between 2 and 3, then instantly jumps to $1/2 + 1/4 + 1/8 = 7/8$ at $x = 3$.

Continuing this pattern, we see the cdf jumps up by smaller and

smaller amounts at each positive integer, and in fact

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1/2 & \text{if } 1 \leq x < 2 \\ 3/4 & \text{if } 2 \leq x < 3 \\ 7/8 & \text{if } 3 \leq x < 4 \\ 15/16 & \text{if } 4 \leq x < 5 \\ \vdots & \\ 2^n - 1/2^n & \text{if } n \leq x < n+1 \\ \vdots & \end{cases}$$

**Exercise 8.4.**
Compute the cdf of the random variable from Example 8.3.

Just as the pmf has some nice properties that hold for every discrete random variable, the cdf also has properties that always hold.

**Theorem 8.3.**
*Let $X$ be any discrete random variable and let $F : \mathbb{R} \to \mathbb{R}$ denote the cdf of $X$. Then we have the following:*

1. *$F$ is an increasing function,*
2. $\lim_{x \to -\infty} F(x) = 0$,
3. $\lim_{x \to \infty} F(x) = 1$, *and*
4. *$F$ is continuous from the right; i.e.,* $F(x) = \lim_{a \to x^+} F(a)$.

**Remark.**
The proof of Theorem 8.3 is a bit technical, so this is just a quick reminder that you can skip over reading the proofs if you feel they're too difficult: you'll never be asked to recite one of these proofs on a quiz or exam. The proofs are provided so that if you're the kind of student that's curious about why these theorems are true, the details are provided if you're willing to wade through them.

In the proof below we use a fact from calculus that you may not be aware of: if a limit exists, it is unique. In particular, to calculate a limit like $\lim_{x\to a} f(x)$, it suffices to consider a sequence of numbers $x_1$, $x_2$, $x_3$, ... which approach $a$ (so, $\lim_{n\to\infty} x_n = a$) and calculate the limit of $f(x)$ along the values in this sequence:

$$\lim_{x\to a} f(x) = \lim_{n\to\infty} f(x_n).$$

This also works for one-sided limits, where for a left-hand limit $\lim_{x\to a^-} f(x)$ we would want a sequence $x_n$ which increases to $a$, and for a right-hand limit $\lim_{x\to a^+} f(x)$ we would want a sequence which decreases to $a$.

The reason we're rewriting our limits in terms of sequences like this is so we can take advantage of Propositions 4.9 and 4.10 which are stated in terms of sequences.

*Proof.*

1. Recall that a function $F : \mathbb{R} \to \mathbb{R}$ is called increasing if for every pair of real numbers $x_1$ and $x_2$ where $x_1 \leq x_2$, we have $F(x_1) \leq F(x_2)$. (In calculus you learned that for differentiable functions this is the same as saying the derivative is non-negative. However, the cdf of a discrete random variable is not differentiable, so we have to resort to the definition of increasing given above.)

   We need to show the cdf $F$ of a discrete random variable $X$ is increasing. Notice that if $x_1 \leq x_2$, then $(-\infty, x_1] \subseteq (-\infty, x_2]$. This means $X^{-1}((-\infty, x_1]) \subseteq X^{-1}((-\infty, x_2])$. By Proposition 4.3,

this means $\Pr(X^{-1}((-\infty, x_1])) \leq \Pr(X^{-1}((-\infty, x_1]))$, but this exactly means $F(x_1) \leq F(x_2)$.

2. By definition,

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} \Pr(X^{-1}((-\infty, x])).$$

Notice that since $x$ decreases to $-\infty$, $X^{-1}((-\infty, x])$ forms a non-increasing sequence of events. By Proposition 4.10 we can write

$$\lim_{x \to -\infty} \Pr(X^{-1}((-\infty, x])) = \Pr\left(\bigcap_{x=1}^{\infty} X^{-1}((-\infty, -x])\right)$$

(We reversed the order above to agree with the way Proposition 4.10 was written.) Now writing

$$\bigcap_{x=1}^{\infty} X^{-1}((-\infty, -x]) = X^{-1}\left(\bigcap_{x=1}^{\infty}(-\infty, -x]\right)$$

and noting $\bigcap_{x=1}^{\infty}(-\infty, -x] = \emptyset$, we see

$$\lim_{x \to -\infty} F(x) = \Pr(\emptyset) = 0.$$

3. The proof of part (3) of the theorem is identical to the proof of part (2), except that we have a non-decreasing sequence and so apply Proposition 4.9 and use that $\Pr(\Omega) = 1$.

4. Since we're taking the limit as $a$ goes to $x$ from the right, the $a$ values are decreasing to $x$. Letting $a_1$, $a_2$, $a_3$, ... be a decreasing sequence of numbers that approaches $x$, we can apply

Proposition 4.10 to write

$$\begin{aligned}\lim_{a\to x^+} F(a) &= \lim_{n\to\infty} F(a_n)\\ &= \lim_{n\to\infty} \Pr(X^{-1}(-\infty, a_n])\\ &= \Pr\left(X^{-1}\left(\bigcap_{n=1}^{\infty}(-\infty, a_n]\right)\right)\\ &= \Pr(X^{-1}((-\infty, x]))\\ &= F(x).\end{aligned}$$

□

What is perhaps more interesting than Theorem 8.3 is that any function satisfying the three conditions described in that theorem is actually the cdf of some random variable. In general that random variable may not be discrete, but the point is that the cdf actually encodes everything about the random variable: if you know the cdf, you know everything there is to know about the random variable. In particular, you can recover the pmf of a discrete random variable from its cdf.

**Proposition 8.4.**
*Suppose $F$ is the cdf of a discrete random variable $X$. Then the pmf $p$ of $X$ is given by*

$$p(x) = F(x) - \lim_{a\to x^-} F(a).$$

**Remark.**
Before proving Proposition 8.4, let's notice that at the points where $F$ is continuous, the above expression is equal to zero since the left- and right-hand limits will equal one another where the function is continuous. Thus all the non-zero values of the pmf must occur at discontinuities of $F(x)$. Since $F$ is continuous from the right, this means the expression above actually measures the "jumps" at these

discontinuities.

*Proof.*
In the expression $\lim_{a \to x^-} F(a)$, let $a_1$, $a_2$, $a_3$, ... be an increasing sequence which converges to $x$. Thus

$$\lim_{a \to x^-} F(a) = \lim_{n \to \infty} F(a_n) = \lim_{n \to \infty} \Pr(X^{-1}((-\infty, a_n])).$$

Since $a_n$ is an increasing sequence, by Proposition 4.9, we have

$$\lim_{n \to \infty} \Pr(X^{-1}((-\infty, a_n])) = \Pr(X^{-1}\left(\bigcup_{n=1}^{\infty}(-\infty, a_n]\right) = \Pr(X^{-1}((-\infty, x))).$$

That is,

$$F(x) - \lim_{a \to x^-} F(a) = \Pr(X^{-1}(-\infty, x]) - \Pr(X^{-1}(-\infty, x)),$$

but this is the probability of $X$ giving some value less-than-or-equal-to $x$ minus the probability of $X$ giving a value strictly less-than $x$, all that's left over is the probability $X$ equals exactly $x$, but this is precisely $p(x)$. □

**Example 8.6.**
Suppose $X$ is a discrete random variable with the following cdf,

$$F(x) = \begin{cases} 0 & \text{if } x < -1 \\ 1/4 & \text{if } -1 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 7/8 & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3 \end{cases}$$

What is the pmf of $X$?

As stated above, the pmf of $X$ will be zero everywhere except at the discontinuities of $F$. The discontinuities of $F$ occur at $-1$, 1, 2, and 3. At these values we apply Proposition 8.4 to calculate $p(x)$:

$$\begin{aligned} p(-1) &= F(-1) - \lim_{x \to -1^-} F(x) = 1/4 - 0 = 1/4 \\ p(1) &= F(1) - \lim_{x \to 1^-} F(x) = 3/4 - 1/4 = 1/2 \\ p(2) &= F(2) - \lim_{x \to 2^-} F(x) = 7/8 - 3/4 = 1/8 \\ p(3) &= F(3) - \lim_{x \to 3^-} F(x) = 1 - 7/8 = 1/8 \end{aligned}$$

Thus the pmf is

$$p(x) = \begin{cases} 1/4 & \text{if } x = -1 \\ 1/2 & \text{if } x = 1 \\ 1/8 & \text{if } x = 2 \\ 1/8 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

**Exercise 8.5.**
Suppose the cdf $F$ of a discrete random variable $X$ is given as follows:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \leq x < 1 \\ 3/5 & \text{if } 1 \leq x < 2 \\ 4/5 & \text{if } 2 \leq x < 3 \\ 9/10 & \text{if } 3 \leq x < 3.5 \\ 1 & \text{if } x \geq 3.5 \end{cases}$$

Compute the pmf $p(x)$ of the random variable.

## 8.3 Expected value

As already mentioned, we often think of random variables as giving us a random real number and forget about the underlying experiment. Sometimes we may want to know what the "average" random number determined by the random variable is. Notice that, depending on the random variable, some real numbers may be much more or less likely than others, and so our notion of average should somehow be aware of these likelihoods. By weighing the numbers in the average by the likelihood of that number, we obtain the *expected value* of the random variable.

If $X$ is a discrete random variable with probability mass function $p(x)$, the ***expected value*** of $X$, denoted $\mathbb{E}[X]$, is the value

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

As $p(x)$ is the probability that the random variable $X$ spits out the value $x$, this is the sum of each real number $x$ times the probability we see that value.

**Example 8.7.**
Consider again the random variable $X$ which counts the number of heads obtained in flipping a fair coin three times. We saw in Example 8.1 the pmf of this random variable was

$$p(x) = \begin{cases} 1/8 & \text{if } x = 0 \\ 3/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 2 \\ 1/8 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

Since "most" real numbers have $p(x) = 0$, these don't contribute to the sum since $x \cdot p(x)$ will be zero for those terms. This means the

expected value can be written as

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\mathbb{R}} x\,p(x)\\ &= 0\cdot {}^1\!/\!_8 + 1\cdot {}^3\!/\!_8 + 2\cdot {}^3\!/\!_8 + 3\cdot {}^1\!/\!_8\\ &= {}^3\!/\!_8 + {}^6\!/\!_8 + {}^3\!/\!_8\\ &= {}^{12}\!/\!_8\\ &= {}^3\!/\!_2\\ &= 1.5\end{aligned}$$

So the "average" output of this random variable is $^3\!/\!_2$.

As the previous example shows, this "average" value of a random variable does not need to be a value that the random variable actually takes on. This might seem odd the first time you see it, so how should you interpret such a value. Looking at the pmf in the example above, notice that one heads and two heads are the most likely scenarios, and these each have the same probabilities. The expected value is right in between these two most common possibilities since $^3\!/\!_2 = 1.5$. If we modified the pmf above so that $X = 1$ was more likely than $X = 2$, this would pull the expected value down closer to 1 to compensate for the more likely $X = 1$.

**Example 8.8.**
Compute the expected value of the random variable with the following pmf:

$$p(x) = \begin{cases} {}^1\!/\!_8 & \text{if } x = 0\\ {}^1\!/\!_2 & \text{if } x = 1\\ {}^1\!/\!_4 & \text{if } x = 2\\ {}^1\!/\!_8 & \text{if } x = 3\\ 0 & \text{otherwise}\end{cases}$$

Computing $\mathbb{E}[X]$ as before we have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x\in\mathbb{R}} x\,p(x)\\ &= 0\cdot 1/8 + 1\cdot 1/2 + 2\cdot 1/4 + 3\cdot 1/8\\ &= 1/2 + 1/2 + 3/8\\ &= 11/8\\ &= 1.375\end{aligned}$$

Notice how the expected value got pulled down closer to 1 when we modified the pmf so that $X = 1$ is more likely than $X = 2$.

**Exercise 8.6.**
Suppose $X$ is the random variable which tells you the value of a roll of a fair, six-sided die. What is the expected value of $X$?

**Exercise 8.7.**
Consider an experiment where again a fair, six-sided die is rolled. However, the sides of this die have values 1, 2, 3, 4, 5, and 100. What is the expected value of a roll of the die?

**Example 8.9.**
The grand prize for the PowerBall lottery in February 2018 was \$203,000,000. According to the PowerBall website, the odds of winning the grand prize are one in 292, 201, 388. If this was the only prize value (so we're assuming no other prizes, just to make the computation simpler), what would the expected value of a \$2 PowerBall ticket be?

Here we have two possibilities, we either win the grand prize or we win nothing. Since we have to spend \$2 to buy a ticket, this means

we either lose two dollars (if we win nothing), or we win \$202,999,998 (if we win the grand prize, minus the cost of the ticket). Multiplying these by the probabilities of each possibility we have

$$\begin{aligned} & 202999998 \cdot \frac{1}{292201388} + (-2) \cdot \frac{292201387}{292201388} \\ \approx & -1.31 \end{aligned}$$

You should interpret the expected value from Example 8.9 as follows: if you were to play the PowerBall lottery over and over and over again, each time buying a two dollar ticket and either winning the grand prize (which is very unlikely) or simply losing your two dollars (much more likely), then on average you will lose \$1.31. That is, you lose your two dollars much more often than you win the two-hundred million dollars. Of course, if you played the lottery enough (meaning almost three-hundred million times) you'd occassionally win, and this balances out all of the times you lost money so that on average you're losing \$1.31 instead of \$2.00.

**Exercise 8.8.**
Compute the expected value of the random variable $X$ with the following pmf:

$$p(x) = \begin{cases} 7/15 & \text{if } x = -4 \\ 2/15 & \text{if } x = -2 \\ 1/5 & \text{if } x = 0 \\ 1/15 & \text{if } x = 1 \\ 2/15 & \text{if } x = 3 \end{cases}$$

## 8.4 Functions of random variables

Before we move on and discuss standard deviation and variance in the next section, let's go ahead and make an observation about random variables which will be helpful later. If $X$ is a random variable defined on a sample space $\Omega$, then $X$ is by definition a function $X : \Omega \to \mathbb{R}$. How, suppose you had another function $f : \mathbb{R} \to \mathbb{R}$ – just a good, ol' fashioned "normal"

function like you're used to from algebra or calculus, maybe something like $f(x) = x^2$, $f(x) = \sin(x)$, or $f(x) = \log(|x| + 1)$. Notice that we could compose $X$ and $f$ to get a new function: we take a point $\omega \in \Omega$, plug it into $X$ to get a real number $X(\omega)$, and then we could take that real number and plug it into $f$ to get another real number, $f(X(\omega))$. Chaining the functions $X$ and $f$ together like this results in a new function which takes inputs in $\Omega$ and ultimately gives you back a real number – that is, this is a new random variable! Technically this random variable should be written as $f \circ X$, but usually we'll just denote it $f(X)$ (note the capital $X$ in the parentheses).

To help this make a little more sense, let's consider a concrete example.

**Example 8.10.**
Suppose $X$ is a discrete random variable with the following pmf:

$$p(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

(Here we take the point of view that the underlying experiment doesn't matter: as long as we know the pmf, we know everything we need to know about the random variable.)

Now suppose $f(x)$ is some "traditional" function, say $f(x) = x^2 - 3x + 2$. Our random variable $X$ above spits out either 1, 2, or 3. Whatever $X$ gives us we'll plug into $f$ to get a new value $f(X)$. Notice if $X = 1$ or $X = 2$, the value of $f(X)$ will be zero (since $1^2 - 3 \cdot 1 + 2 = 2^2 - 3 \cdot 2 + 2 = 0$). If $X = 3$, then the value of $f(X)$ will be 2.

So, $f(X)$ is a random variable that spits out either 0 or 2. Notice the probability $f(X) = 0$ would be the probability $X = 1$ or $X = 2$, since these are the only outputs of $X$ which give $f(X)$ equal to zero. From the pmf above, this means the probability $f(X) = 0$ is $1/2 + 1/3 = 5/6$. Similarly, the probability $f(X) = 2$ would be the probability $X = 3$ and so must equal $1/6$. That is, the pmf of the random variable $f(X)$, which we'll denote $p_f(x)$ to distinguish it from

the pmf of $X$ above, is

$$p_f(x) = \begin{cases} 5/6 & \text{if } x = 0 \\ 1/6 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Once we have a random variable, we may want to know what its expected value is.

**Example 8.11.**
What is the expected value of the random variable $f(X)$ from Example 8.10?

Since we computed the pmf of $f(X)$ in Example 8.10, we can easily compute the expected value:

$$\mathbb{E}[f(X)] = 0 \cdot 5/6 + 2 \cdot 1/6 = 1/3.$$

Based on what we've done thus far, you might reasonably assume that to find the expected value of $f(X)$ you'd need to determine the pmf of $f(X)$. In the example above this wasn't too hard, but in more involved examples this can actually be really difficult. In particular, to find the pmf of $f(X)$, for each possible output $x$ we'd have to find the probability $f(X)$ equals $x$ (by definition of the pmf). In general, to do this we would have to find all possible outputs of $X$ which when plugged into $f$ give us $x$, and this is usually very hard.

In particular, if $p(x)$ is the pmf of the original random variable $X$ and we wanted to find the pmf $p_f(x)$ of the "new" random variable $f(X)$ we get after composing with $f$, we would have to compute the following:

$$p_f(x) = \Pr(f(X) = x) = \Pr(X \in f^{-1}(x)) = \Pr(X^{-1}(f^{-1}(x)).$$

The issue is that $f^{-1}(x)$ could consist of several possible values. Say, for the sake of example, $f^{-1}(x) = \{w_1, w_2, ...\}$ for some (possibly infinite) set of values of $w_i$. Now we still have to find the values of $\Omega$ which $X$ maps to each $w_i$:

$$X^{-1}(f^{-1}(x)) = X^{-1}(\{w_1, w_2, ...\}) = \bigcup_{w \in f^{-1}(x)} X^{-1}(w)$$

Now supposing this collection of $w$ values is not "too infinite"[1], we have that the pmf of $f(X)$ is

$$p_f(x) = \sum_{y \in f^{-1}(x)} p(y).$$

This might lead you to believe that finding $\mathbb{E}[f(X)]$ is going to be hard in general since it's usually going to be difficult to compute $p_f(x)$ using the expression above. Luckily, however, there's a trick that let's us sidestep having to compute the pmf of $f(X)$.

**Theorem 8.5.**
*If $X$ is a discrete random variable with pmf $p(x)$ and $f$ is any function $f : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[f(X)] = \sum_{x \in \mathbb{R}} f(x)p(x)$$

Before giving the proof of Theorem 8.5, let's notice why this is such a nice theorem to have. As we indicated before the theorem, finding the pmf of a function $f(X)$ of a random variable $X$ is generally going to be difficult. Theorem 8.5 says that we don't have to do that hard work, however. In fact, we can compute $\mathbb{E}[f(X)]$ with a formula very similar to the formula for $\mathbb{E}[X]$: all we have to do is change the factor of $x$ in $\sum_{x \in \mathbb{R}} x\, p(x)$ to an $f(x)$.

*Proof of Theorem 8.5.*
Since $X$ is a discrete random variable, so is $f(X)$, and so $f(X)$ must have some pmf which we'll denote $p_f(x)$. As noted above, this $p_f(x)$ is given by the formula

$$p_f(x) = \sum_{y \in f^{-1}(x)} p(y)$$

where $p$ is the pmf of $X$. Plugging this into the "usual" equation for

[1] In particular, assuming for each $x \in \mathbb{R}$, $X^{-1}(w)$ is a countable set. This is a technical assumption we need; don't worry about it if it doesn't make sense to you.

the expected value gives us

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_{x\in\mathbb{R}} x\, p_f(x) \\ &= \sum_{x\in\mathbb{R}} x \sum_{y\in f^{-1}(x)} p(y) \\ &= \sum_{x\in\mathbb{R}} \sum_{y\in f^{-1}(x)} xp(y)\end{aligned}$$

Now we do something very simple, but a little bit clever. Since $y$ is in the preimage of $x$, this means $f(y) = x$ and so we can rewrite the $x$ above as $f(y)$ to obtain

$$\mathbb{E}[f(X)] = \sum_{x\in\mathbb{R}} \sum_{y\in f^{-1}(x)} f(y)p(y).$$

Now simply notice that each $y$ occurs exactly once in the sums above: in particular, $y$ occurs once in the inner sum when $x$ in the outer sum equals $f(y)$. That is, the above double sum can be written more simply as a single sum,

$$\mathbb{E}[f(X)] = \sum_{y\in\mathbb{R}} f(y)p(y).$$

Of course, whether we call the variable $y$ or $x$ or ☺ or anything else doesn't really matter, and hence writing the $y$ above as $x$ gives the result. □

**Example 8.12.**
To verify the formula from Theorem 8.5 above works, let's recompute the expected value from Example 8.11 using the expression from Theorem 8.5.

Recall $X$ was the discrete random variable with pmf

$$p(x) = \begin{cases} 1/2 & \text{if } x = 1 \\ 1/3 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

and $f(x) = x^2 - 3x + 2$. Now we compute

$$\begin{aligned} \mathbb{E}[f(X)] &= \sum_{x \in \mathbb{R}} f(x)p(x) \\ &= f(1)\,p(1) + f(2)\,p(2) + f(3)\,p(3) \\ &= 0 \cdot 1/2 + 0 \cdot 1/3 + 2 \cdot 1/6 \\ &= 1/3 \end{aligned}$$

which was the same as the expected value computed in Example 8.11 using the pmf of $f(X)$.

For some very simple types of functions $f(x)$ we can actually give an even simpler shortcut for computing $\mathbb{E}[f(X)]$. If $f(x)$ is a function whose graph is a line (such functions are called *affine functions*[2]), i.e., if $f(x)$ has the form

$$f(x) = mx + b$$

for some constants $m$ and $b$, then

$$\mathbb{E}[f(X)] = \mathbb{E}[mX + b] = m\mathbb{E}[X] + b.$$

That is, we can split up expected values by breaking up sums and pulling out constants. If you write out the for $\mathbb{E}[f(X)]$ using Theorem 8.5 and perform the tiniest bit of algebra and one earlier property of random variables, you can easily justify the expression above, and so we leave the proof of this fact as an exercise.

[2]You might be tempted to call such an $f(x)$ a *linear function* since its graph is a line, but in some areas of math, such as linear algebra, the expression "linear function" would mean something more specific, so it's best if we avoid calling this a linear function.

**Exercise 8.9.**
Show that for any discrete random variable $X$ and any constants $m$ and $b$,

$$\mathbb{E}[mX + b] = m\mathbb{E}[X] + b.$$

(Hint: Use Theorem 8.5 and Corollary 8.2.)

## 8.5 Variance and standard deviation

The expected value of $X$ tells us what the "center" of the values of $X$ is; one way to think of $\mathbb{E}[X]$ is that it is the center of mass of points spread out on the real line where each point has weight $p(x)$. However, the actual values that $X$ takes on can be very different from this "center of mass," as the next example illustrates.

**Example 8.13.**
Consider random variables $X$, $Y$, and $Z$ with the following pmf's:

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = -1 \\ 1/2 & \text{if } x = 1 \end{cases}$$

$$p_Y(x) = \begin{cases} 1/4 & \text{if } x = -2 \\ 1/4 & \text{if } x = -1 \\ 1/2 & \text{if } x = 3/2 \end{cases}$$

$$p_Z(x) = \begin{cases} 1/5 & \text{if } x = -15 \\ 1/10 & \text{if } x = -5 \\ 1/2 & \text{if } x = -3 \\ 1/10 & \text{if } x = 20 \\ 1/10 & \text{if } x = 30 \end{cases}$$

A simple calculation shows that each random variable has the same expected value of 0.

Even though each random variable in Example 8.13 has the same expected value (the same "center of mass"), the actual values the random variable takes on are distributed around that expected very differently. We would like to have some way of measuring how far the values the random variables takes on differ, on average, from the expected value. There are two related notions of this, the variance and the standard deviation.

Let's notice that we're trying to find how far the values of $X$ are from the expected value $\mathbb{E}[X]$. One natural thing to consider would simply be the difference, $X - \mathbb{E}[X]$. Note that this quantity is actually a new random variable. To see this, let's make one notational simplification: the value $\mathbb{E}[X]$, whatever it happens to be, is just some number, and let's call that number $\mu$ for the moment. So, we're interested in $X - \mu$. Notice this is just the composition $f(X)$ where $f(x) = x - \mu$, and so by the discussion in the previous section we have a random variable.

Now, we want to know the average of value $X - \mu$, so we might try to take the expected value of this new random variable and compute $\mathbb{E}[X - \mu]$. By Exercise 8.9, we can easily compute this quantity:

$$\mathbb{E}[X - \mu] = \mathbb{E}[X] - \mu.$$

Keeping in mind $\mu$ is really just shorthand for $\mathbb{E}[X]$, however, we see this quantity will always be zero, and so it's not very helpful for us.

The issue is that $X - \mu$ will sometimes be a little bigger than $\mu$ (so $X - \mu > 0$) and sometimes a little bit smaller than $\mu$ (so $X - \mu < 0$). Since $\mu$ is the "center" of the data, these positives and negatives perfectly balance out and always give zero. We can easily fix this by squaring $X - \mu$, since this forces everything to be positive. That is, we consider the expected value of $(X - \mu)^2$, and this is what we define the *variance* of $X$ to be.

The ***variance*** of a random variable $X$ is the expected value of $(X - \mathbb{E}[X])^2$, and denote this quantity $\mathrm{Var}(X)$:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Since it's easy to get confused with lots of $\mathbb{E}$'s floating around, we often write $\mu$ for $\mathbb{E}[X]$ so that we can write the variance as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2].$$

By Theorem 8.5, we can compute the variance using the formula

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] = \sum_{x \in \mathbb{R}} (x - \mu)^2 p(x).$$

So, computing the variance is easy to do if we have the pmf of $X$. Note, though, that we need to compute $\mu = \mathbb{E}[X]$ before we can compute $\text{Var}(X)$.

As a few first easy examples, let's go ahead and compute the variance of the random variables $X$, $Y$, and $Z$ from Example 8.13 since we've already computed $\mu = 0$ for each random variable.

**Example 8.14.**
The variance of the random variable $X$ from Example 8.13 is

$$\begin{aligned}\text{Var}(X) &= \sum_{x\in\mathbb{R}}(x-0)^2 p_X(x)\\ &= (-1)^2\cdot 1/2 + 1^2\cdot 1/2\\ &= 1/2 + 1/2\\ &= 1\end{aligned}$$

This means the average square of the difference between an output of $X$ and the expected value 0 is 1, which seems completely obvious from the pmf.

The variance of the random variable $Y$ from Example 8.13 is

$$\begin{aligned}\text{Var}(Y) &= \sum_{x\in\mathbb{R}}(x-0)^2 p_Y(x)\\ &= (-2)^2\cdot 1/4 + (-1)^2\cdot 1/4 + (3/2)^2\cdot 1/2\\ &= 1 + 1/4 + 9/8\\ &= 19/8\end{aligned}$$

So the square of the distance between $X$ and 0 is, on average, $19/8$.

The variance of the random variable $Z$ from Example 8.13 is

$$\begin{aligned}\text{Var}(Z) &= (-15)^2\cdot 1/5 + (-5)^2\cdot 1/10 + (-3)^2\cdot 1/2 + 20^2\cdot 1/10 + 30^2\cdot 1/10\\ &= 182\end{aligned}$$

Now, let's notice something about the variances calculated above. Even though you may not have a lot of intuition about what these numbers are, you should notice that the more spread out our data was, the bigger the variance was. To put this in perspective, let's consider two more simple examples.

**Example 8.15.**
Consider the random variable $X$ with pmf

$$p(x) = \begin{cases} 2/3 & \text{if } x = 0.9 \\ 1/3 & \text{if } x = 1.2. \end{cases}$$

To compute the variance of this random variable, we first need to know its expected value,

$$\mathbb{E}[X] = 0.9 \cdot 2/3 + 1.2 \cdot 1/3 = 1.$$

The variance can now be computed as

$$\mathrm{Var}(X) = (0.9 - 1)^2 \cdot 2/3 + (1.2 - 1)^2 \cdot 1/3 = 0.02$$

**Example 8.16.**
Consider the random variable $X$ with pmf

$$p(x) = \begin{cases} 2/3 & \text{if } x = -99 \\ 1/3 & \text{if } x = 201 \end{cases}$$

To compute the variance, we need the expected value,

$$\mathbb{E}[X] = -99 \cdot 2/3 + 201 \cdot 1/3 = 1.$$

The variance is thus

$$\mathrm{Var}(X) = (-99 - 1)^2 \cdot 2/3 + (201 - 1)^2 \cdot 1/3 = 20000.$$

Again, the random variables in Examples 8.15 and 8.16 had the same expected value, but in Example 8.15 the values of $X$ were very close to that expected value, while in Example 8.16 the values were very far away. Correspondingly, the variance of Example 8.15 was very small, and the variance of Example 8.16 was very large. This is the whole point: the variance gives us a way of comparing two random variables to see if their

outputs are tightly packed together near the expected value, or if they're more spread out.

The variance is often given by another equivalent formula, which we can get by simply doing some algebra and applying some basic properties of expected values and pmf's.

**Lemma 8.6.**
*If $X$ is a discrete random variable with pmf $p(x)$ and expected value $\mathbb{E}[X] = \mu$, then the variance of $X$ is equal to*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2.$$

*Proof.*

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_{x\in\mathbb{R}} (x-\mu)^2 p(x) \\
&= \sum_{x\in\mathbb{R}} \left(x^2 - 2x\mu + \mu^2\right) p(x) \\
&= \sum_{x\in\mathbb{R}} \left(x^2 p(x) - 2x\mu p(x) + \mu^2\right) p(x) \\
&= \sum_{x\in\mathbb{R}} x^2 p(x) - \sum_{x\in\mathbb{R}} 2x\mu p(x) + \sum_{x\in\mathbb{R}} \mu^2 p(x) \\
&= \sum_{x\in\mathbb{R}} x^2 p(x) - 2\mu \sum_{x\in\mathbb{R}} x p(x) + \mu^2 \sum_{x\in\mathbb{R}} p(x) \\
&= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\
&= \mathbb{E}[X^2] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}[X^2] - \mu^2
\end{aligned}$$

□

Just to convince ourselves both formulas for the variance give the same result, let's compute the variance of a discrete random variable with each formula.

**Example 8.17.**
Consider the random variable $X$ with pmf

$$p(x) = \begin{cases} 1/10 & \text{if } x = 1 \\ 1/4 & \text{if } x = 4 \\ 1/2 & \text{if } x = 8 \\ 3/20 & \text{if } x = 10 \end{cases}$$

Compute the variance $\mathrm{Var}(X)$ using both formulas,

$$\mathbb{E}[(X-\mu)^2] \quad \text{and} \quad \mathbb{E}[X^2] - \mu^2.$$

For each formula we need to first compute $\mu = \mathbb{E}[X]$:

$$\mu = \mathbb{E}[X] = 1 \cdot 1/10 + 4 \cdot 1/4 + 8 \cdot 1/2 + 10 \cdot 3/20 = 33/5.$$

Now we compute the variance with each formula:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[(X-\mu)^2] \\ &= \mathbb{E}[(X - 33/5)^2] \\ &= (1 - 33/5)^2 \cdot 1/10 + (4 - 33/5)^2 \cdot 1/4 + (8 - 33/5)^2 \cdot 1/2 + (10 - 33/5)^2 \cdot 3/20 \\ &= 377/50. \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - \mu^2 \\ &= \left(1^2 \cdot 1/10 + 4^2 \cdot 1/4 + 8^2 \cdot 1/2 + 10^2 \cdot 3/20\right) - (33/5)^2 \\ &= 511/10 - 1089/25 \\ &= 377/50. \end{aligned}$$

Though the numbers in the arithmetic above are ugly, we see that the two formulas give us the same value in the end.

Since the variance is defined as an expectation (using our earlier formula), we can adapt the formula for expected value of a function of a random variable,

$$\mathbb{E}[f(X)] = \sum_{x \in \mathbb{R}} f(x)p(x),$$

to get a formula for the variance of $f(X)$:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[(f(X) - \mathbb{E}[f(X)])^2] \\ &= \sum_{x\in\mathbb{R}} [f(x) - \mathbb{E}[f(X)]]^2 \, p(x).\end{aligned}$$

This formula is admittedly not as nice as our formula for the expectation of $f(X)$, although if our function $f$ is simple enough, then the formula above will simplify nicely.

**Proposition 8.7.**
*If $X$ is a discrete random variable with pmf $p(x)$, then for any constants $m$ and $b$,*

$$\text{Var}(mX + b) = m^2\text{Var}(X).$$

Notice that in Proposition 8.7 we are composing the random variable $X$ with the function $f(x) = mx + b$. The proposition says two things about such a composition: if we slide all of the values of $X$ over by a constant, it doesn't chance the variance; and if we multiply all the values of $X$ by a constant, the variance changes by the square of that constant.

*Proof of Proposition 8.7.*

$$\begin{aligned}\text{Var}(mX + b) &= \sum_{x\in\mathbb{R}} (mx + b - \mathbb{E}[mX + b])^2 \, p(x) \\ &= \sum_{x\in\mathbb{R}} (mx + b - m\mathbb{E}[X] - b)^2 \, p(x) \\ &= \sum_{x\in\mathbb{R}} m^2(x - \mathbb{E}[X])^2 p(x) \\ &= m^2 \sum_{x\in\mathbb{R}} (x - \mathbb{E}[X])^2 p(x) \\ &= m^2\mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= m^2\text{Var}(X)\end{aligned}$$

□

**Example 8.18.**
If $X$ is the discrete random variable from Example 8.17, then

$$\mathrm{Var}(\sqrt{50}X + \pi^{\sqrt{e}}) = 377.$$

In computing variance, recall that we had to introduce a square in order to avoid having all of the terms cancel. Because of this square, the numbers that appear in variance calculations are often much bigger than the kinds of numbers you might intuitively guess measure the spread of the values of $X$. To compensate for this we can do the obvious thing: let's take the square root. That is, we often consider the quantity $\sqrt{\mathrm{Var}(X)}$ instead of $\mathrm{Var}(X)$ directly. This is called the ***standard deviation*** of $X$, and in many real-world applications people tend to discuss the standard deviation more than the variance. Even though variance and standard deviation are essentially the same thing, because of the square root the numbers that appear as standard deviations are usually more intuitive than the numbers that appear in variance calculations – but all you're doing is calculating the variance first and then taking the square root, so there's not really any new math in computing standard deviations.

We often use the lowercase Greek letter sigma, $\sigma$, to denote standard deviations and correspondingly use $\sigma^2$ to denote the variance. If we are considering several random variables at once, we may want to keep track of which standard deviation (or variance) is associated with which variable, and we may use subscripts to do this. E.g., $\sigma_X$ is the standard deviation of a random variable $X$, and $\sigma_Y^2$ is the variance of a random variable $Y$. When there's no risk for confusion (i.e., when we're only discussing one random variable), we often don't bother with the subscripts.

**Exercise 8.10.**
Show that if $X$ is a discrete random variable and if $m$ and $b$ are any two constants, then

$$\sigma_{mX+b} = |m|\sigma_X.$$

## 8.6 Practice problems

**Problem 8.1.**
What value of $k$ makes the function below the cdf of a discrete random variable?

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ k & \text{if } 0 \leq x < 1 \\ k + k/3 & \text{if } 1 \leq x < 2 \\ k + k/3 + k/9 & \text{if } 2 \leq x < 3 \\ k + k/3 + k/9 + k/27 & \text{if } 3 \leq x < 4 \\ \vdots & \\ \sum_{j=0}^{n-1} k/3^j & \text{if } n-1 \leq x < n \text{ where } n \text{ is a positive integer} \\ \vdots & \end{cases}$$

**Problem 8.2.**
Suppose that $X$ is a random variable with expected value $\mu$. Compute $\mathbb{E}[X - \mu]$.

**Problem 8.3.**
Suppose $X$ is a discrete random variable with the following probability mass function:

$$p(x) = \begin{cases} 1/10 & \text{if } x = 1 \\ 1/5 & \text{if } x = 2 \\ 2/5 & \text{if } x = 3 \\ 3/10 & \text{if } x = 4 \\ 0 & \text{otherwise} \end{cases}$$

Let $f(x) = x^2 - 5x + 4$. Compute the expected value of $f(X)$.

**Problem 8.4.**
Suppose $X$ is a random variable with the following cumulative distribution function. What is the expected value of $X$?

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/8 & \text{if } 1 \leq x < 3 \\ 3/4 & \text{if } 3 \leq x < 5 \\ 1 & \text{if } x \geq 5 \end{cases}$$

**Problem 8.5.**
Suppose $X$ is a random variable with the following cumulative distribution function. What is the probability mass function of $X$?

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/8 & \text{if } 1 \leq x < 3 \\ 3/4 & \text{if } 3 \leq x < 5 \\ 1 & \text{if } x \geq 5 \end{cases}$$

**Problem 8.6.**
Compute the variance in the random variable obtained by summing the rolls of two "normal," fair six-sided dice.

**Problem 8.7.**
Consider rolling two six-sided dice which are special in the following way: one die has three red sides, two blue sides, and one green side; the other die has two red sides, two blue sides, and two green sides. The two dice are rolled and then a numerical value is associated to the roll in the following way: if the same color is rolled on both dice we assign the value 10; if one die rolls red and one rolls green we assign 8; if one die rolls red and one rolls blue we assign 4; and if one die rolls blue and one rolls green, we assign 3. In this way we have a random variable assigned to the roll of the two dice.

(a) Compute the pmf of this random variable.

(b) Compute the expected value of this random variable.

**Problem 8.8.**
Suppose a car insurance company divides claims from automobile accidents into four categories: trivial claims where the damage incurred in the accident is \$0; minor claims where the damage incurred is \$1000; moderate claims where the damage is \$5000; and serious claims where the damage is \$10,000. Suppose also that 80% of claims are trivial, 10% of claims are minor, 8% of claims are moderate, and 2% of claims are serious. If each customer has a \$500 deductible, what premium should the company charge if it wants to average \$100 in profit per customer?

**Problem 8.9.**
Compute the expected value of the random variable $X$ which has the following pdf:

$$p(x) = \begin{cases} 1/2^n & \text{if } x = 2^n \text{ for some positive integer } n \\ 0 & \text{otherwise} \end{cases}$$

**Problem 8.10.**
Suppose that a certain toll bridge charges \$1 for each car and \$2.50 for each truck that passes over the bridge. Suppose also that 60% of the vehicles travelling over the bridge are cars. If twenty-five vehicles cross the bridge during some given interval of time, what is the expected revenue of the toll?

# 9

# *Families of Discrete Random Variables*

*Ein Mathematiker, der nicht etwas Poet ist, wird nimmer ein vollkommener Mathematiker sein.*
A mathematician who is not something of a poet will never be a good mathematician.

Karl Weierstrass

Many, though not all, of the discrete random variables we often care about in "the real world" are members of one of several families of random variables. By a "family" of random variables here, what we mean is that the random variables have *almost* the same pmf, but with some minor differences that depend on a parameter we must specify. I.e., the random variables in a given family all have the same pmf if we write part of the pmf as a variable, and the different members of these families correspond to different values of that variable.

If this all seems a little strange, don't worry about it right now: the ideas will become clearer after we've seen a few examples.

## 9.1 Bernoulli

The simplest family of discrete random variables are the Bernoulli random variables, named after Jacob Bernoulli, a Swiss mathematician that studied games of chance in the 17th century.

**Remark.**
Two fun facts: For a long time I assumed that the Bernoulli the Bernoulli random variables were named after was the same as the Bernoulli as the Bernoulli principle in physics (the phenomenon that pressure in a fluid decreases as its speed increases – the principle that allows airplanes to fly). These are in fact, different people – Daniel Bernoulli was the physicists. Apparently there was actually a whole family of Bernoullis, mathematicians, physicists, and engineers that

were all related to each other.

The other fun fact is that the mathematical constant $e$ is named after another Swiss mathematician, Leonhard Euler, although Jacob Bernoulli was the first person to study this number, which he realized popped up in the limit when you look at interest that's compounded over smaller and smaller intervals of time.

A ***Bernoulli random variable*** is any random variable $X$ that only takes on one of two values, 0 or 1. The only thing that distinguishes one Bernoulli random variable from another is the probability of 0 and of 1.

For example, the random variable $X$ with the following pmf,

$$p_X(x) = \begin{cases} 1/2 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

is a Bernoulli random variable. So is the random variable $Y$ with pmf

$$p_Y(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 3/2 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

and so is the random variable $Z$ with pmf

$$p_Z(x) = \begin{cases} 3/10 & \text{if } x = 0 \\ 7/10 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

In each case our random variable can only take on the values 0 or 1, so these are the only numbers the pmf gives a non-zero probability to. The only difference between these random variables is the probability of 0 and 1. In fact, we only need to know one of these probabilities, since probabilities have to add up to 1. In particular, if the probability a Bernoulli random variable $X$ equals 1 is $p$, for some number $0 \leq p \leq 1$, then the probability $X$ equals 0 must be $1 - p$.

That is, if we $X$ is a Bernoulli random variable and we also know this value of $p$, we know the pmf of $X$ must be

$$p_X(x) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

We denote that $X$ is this Bernoulli random variable with parameter $p$ by writing $X \sim \text{Bernoulli}(p)$.

**Remark.**
Unfortunately we're now using the letter $p$ to mean two different things: sometimes it means the pmf, and sometimes it means the parameter above. It will usually be clear from context when you see a $p$ whether it refers to a pmf of a parameter, but if we want to make things crystal clear we may write $p_X$ instead of simply $p$ when we refer to the pmf of a random variable $X$.

Since the value of the parameter $p$ tells us everything we need to know about a Bernoulli random variable, we should be able to express the expected value and variance of $X \sim \text{Bernoulli}(p)$ in terms of $p$.

**Proposition 9.1.**
*If* $X \sim \text{Bernoulli}(p)$, *then* $\mathbb{E}[X] = p$ *and* $\text{Var}(X) = p(1-p)$.

*Proof.*
We simply compute the expected value and variance using our formulas from the previous chapter.

$$\mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p.$$

Now that we know $\mathbb{E}[X]$ is $p$, we can compute the variance:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - p^2 \\
&= \left(0^2 \cdot (1-p) + 1^2 \cdot p\right) - p^2 \\
&= p - p^2 \\
&= p(1-p).
\end{aligned}$$

□

In some sense, the Bernoulli random variables are the simplest (and most boring) random variable imaginable, so you may wonder why we even bother discussing them. In many applications we only care about whether an outcome of an experiment occurs or not, and this is easily translated into a Bernoulli random variable if we agree 1 means the outcome occurred and 0 means the desired outcome did not occur.

**Example 9.1.**
Consider the following silly game: A normal, six-sided die is rolled. If the value that appears on the die is 3 or higher, you win \$1, but if the value that appears is 2 or lower, you win nothing. If you were to play this game many times, on average how much would you win per game? What is the standard deviation in your winnings?

This is of course a Bernoulli random variable where success (winning one dollar) corresponds to rolling 3, 4, 5, or 6 on the die, and failure (winning nothing) corresponds to rolling 1 or 2. The probability of success is $4/6 = 2/3$. That is, the random variable indicating whether we win or lose the game is $X \sim \text{Bernoulli}(2/3)$.

The average winnings per game, if we were to play several games, is the expected value. By the formula in the proposition above, the average winnings per game is $2/3$ of a dollar, or (approximately) 66¢. The variance is $p(1-p) = 2/3 \cdot 1/3 = 2/9$, so the standard deviation $\sqrt{2}/3$ of a dollar, which is about 47¢.

We'll see later that many more complicated random variables can be broken up into Bernoulli random variables. For example, the next family of random variables we'll discuss, binomial random variables, can be thought of as sums of Bernoulli random variables. Once we realize this, we can often derive properties of these complicated random variables by thinking of them in terms of Bernoulli random variables. This is a little ways from where we are right now, but that's where we're ultimately heading.

**Exercise 9.1.**

Suppose $X \sim \text{Bernoulli}(p)$. Write down the cdf $F$ of $X$.

## 9.2 Binomial

The next family of random variables we'll discuss are the binomial random variables. A ***binomial random variable*** is a discrete random variable that counts the number of "successes" in a fixed number of independent trials where each trial is classified as a success or failure, and each trial has the same probability of success.

We have actually already seen an example of a binomial random variable earlier in these notes, even though we didn't refer to it as "binomial" at the time. In Example 7.1 we considered flipping three coins and counted the number of heads. If we think of the heads as successes and tails as failures, then this means our random variable counting the number of heads must have been a binomial random variable.

Notice that when discussing a binomial random variable there are two pieces of information we need: the number of trials, and the probability of success on each trial. In our earlier example counting the number of heads when flipping three coins, the number of trials is 3 and the probability of success (heads) is $1/2$. When talking about a binomial random variable in the abstract, we'll often refer to the number of trials as $n$ and the probability of success in each trial as $p$. When then write $X \sim \text{Binomial}(n, p)$ to indicate $X$ is a binomial random variable where we count the number of success among $n$ trials, where the probability of success is $p$. The random variable counting the number of heads in three flips, for instance, is $X \sim \text{Binomial}(3, 1/2)$.

Given this description of a binomial random variable $X \sim \text{Binomial}(n, p)$, we can determine what the pmf $p_X$ of $X$ must be. Of course, we can't have fewer than zero success and we can't have more than $n$ success since there are only $n$ trials. So suppose that $x$ is an integer between 0 and $n$. What is the probability there are exactly $x$ success in our $n$ trials?

If there are $x$ successes among our $n$ trials, then we must choose which of the $n$ trials were the successes; i.e., we must choose $x$ of the $n$ trials to be successes, and the remaining trials will be failures. There are $\binom{n}{x}$ ways we can make this choice. For each choice we need to get success $x$ times and failure $n - x$ times; since the probability of each success is $p$ and the probability of failure is $(1 - p)$, the probability of $x$ successes and $n - x$

failures is $p^x(1-p)^{n-x}$, but we have to add up all the ways we could get those $x$ successes.

Putting all of this together, for an integer $0 \leq x \leq n$, the probability of $x$ successes is

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

Thus the pmf of $X \sim \text{Bernoulli}(n, p)$ is

$$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x \text{ is an integer, } 0 \leq x \leq n \\ 0 & \text{otherwise} \end{cases}$$

**Remark.**
Notice that a binomial random variable with one trial is the same as a Bernoulli random variable: $\text{Binomial}(1, p) = \text{Bernoulli}(p)$.

**Example 9.2.**
Suppose the probability of a given basketball player scoring on a free throw is 0.7. If the player makes free throws, what is the probability they score on exactly eight of those shots? What's the probability they score on *at least* eight shots?

Here we have a random variable $X$ counting the number of shots scored when ten shots are made, where each shot is scores with probability 0.7. This is a binomial random variable with $n = 10$ and $p = 0.7$: $X \sim \text{Bernoulli}(10, 0.7)$. From the formula for the pmf of a binomial above, the probability of scoring exactly eight of the ten free throws is thus

$$\begin{aligned} \Pr(X = 8) &= p_X(8) \\ &= \binom{10}{8} \cdot (0.7)^8 \cdot (0.2)^2 \\ &\approx 45 \cdot 0.0576 \cdot 0.09 \\ &\approx 0.2335 \end{aligned}$$

So there is about a 23.35% chance of scoring exactly eight shots.

To compute the probability of scoring on *at least* eight shots, we have to consider the chance that the player scores on eight shots, nine shots, and ten shots. Plugging these into the pmf for $X$ we have

$$\begin{aligned}
\Pr(X \geq 8) &= \Pr(X = 8 \text{ or } X = 9 \text{ or } X = 10) \\
&= \Pr(X = 8) + \Pr(X = 9) + \Pr(X = 10) \\
&= p_X(8) + p_X(9) + p_X(10) \\
&= \binom{10}{8} \cdot 0.7^8 \cdot 0.3^2 + \binom{10}{9} \cdot 0.7^9 \cdot 0.3^1 + \binom{10}{10} \cdot 0.7^{10} \cdot 0.3^0 \\
&\approx 0.2335 + 0.1211 + 0.0285 \\
&\approx 0.3831
\end{aligned}$$

With any random variable we often want to know the expected value and the variance of that random variable. If $X \sim \text{Binomial}(n, p)$, then the pmf of $X$ depends on these parameters $n$ and $p$, and so it shouldn't be too surprising that we ought to be able to find formulas for the expectation and variance in terms of $n$ and $p$.

**Proposition 9.2.**
*If* $X \sim \text{Binomial}(n, p)$, *then* $\mathbb{E}[X] = np$.

*Proof.*
This is just a calculation using our formulas for expectation and variance, although there are a few bits of algebraic trickery in the computation. We begin by writing out the definition of expected value, then writing out the definition of $\binom{n}{x}$ in terms of factorials and do one

simple cancellation to obtain the following:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathbb{R}} x p_X(x) \\
&= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{x \cdot (x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x}
\end{aligned}$$

Note in the third line above we switched our sum from starting at $x = 0$ to $x = 1$, since the term corresponding to $x = 0$ gets multiplied by zero. We then cancelled the $x$ above with one of the $x$ factors in $x! = x \cdot (x-1)!$.

Now we will write $n!$ as $n \cdot (n-1)!$, $p^x$ as $p \cdot p^{x-1}$, and we will write $n-x$ (which appears twice, once in $(n-x)!$ and once in $(1-p)^{n-x}$) as $n-1-(x-1)$. Plugging all of this into the last line of our computation above gives us

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n \cdot (n-1)!}{(x-1)!(n-1-(x-1))!} p \cdot p^{x-1} \cdot (1-p)^{n-1-(x-1)}.
\end{aligned}$$

This looks complicated, but we will see in a moment that things will simplify nicely. First we distribute out any constants that don't de-

pend on $x$ to write

$$\begin{aligned}&\sum_{x=1}^{n} \frac{n \cdot (n-1)!}{(x-1)!(n-1-(x-1))!} p \cdot p^{x-1} \cdot (1-p)^{n-1-(x-1)} \\ =&np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-1-(x-1))!} p^{x-1} \cdot (1-p)^{n-1-(x-1)}.\end{aligned}$$

Notice that we never have any isolated $x$'s in the terms above: all $x$'s occur as $x-1$. If we introduce a new variable $w$ and set $w = x-1$, then we can rewrite the sum above in terms of $w$ and this sum will start at $w = 0$ since $w = x-1$ and $x$ starts at 0. Since $x$ ends at $n$, $w$ will end at $n-1$:

$$\begin{aligned}&np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-1-(x-1))!} p^{x-1} \cdot (1-p)^{n-1-(x-1)} \\ =&np \sum_{w=0}^{n-1} \frac{(n-1)!}{w!(n-1-w)!} p^{w} \cdot (1-p)^{n-1-w}\end{aligned}$$

Let's also introduce a variable $m = n-1$, so all the $n-1$ terms in our sum can be replaced by $m$,

$$\begin{aligned}&np \sum_{w=0}^{n-1} \frac{(n-1)!}{w!(n-1-w)!} p^{w} \cdot (1-p)^{n-1-w} \\ =&np \sum_{w=0}^{m} \frac{m!}{m!(m-w)!} p^{w} \cdot (1-p)^{m-w}.\end{aligned}$$

Now recall the binomial theorem from basic algebraic which states for any two real numbers $a$ and $b$,

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}.$$

By the binomial theorem, the sum in our expression above can be rewritten as

$$\sum_{w=0}^{m} \frac{m!}{m!(m-w)!} p^{w} \cdot (1-p)^{m-w} = (p + (1-p))^m = 1^m = 1.$$

Plugging this into the above we have

$$\mathbb{E}[X] = np.$$

□

**Proposition 9.3.**
*If* $X \sim \text{Binomial}(n, p)$, *then* $\text{Var}(X) = np(1-p)$.

*Proof.*
The proof of this is similar to the proof for the expected value of a binomial random variable, except there's one little trick that's required. Recall that $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and we've already computed $\mathbb{E}[X] = np$. So all we need to do is to compute $\mathbb{E}[X^2]$, but in order to do this we are going to write $X^2$ as $X(X-1)+X$. This, obviously, is a function of $X$, and so we can compute

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[X(X-1)+X] \\
&= \sum_{x\in\mathbb{R}} (x(x-1)+x)p_X(x) \\
&= \sum_{x=0}^{n} (x(x-1)+x)\binom{n}{x}p^x(1-p)^{n-x} \\
&= \sum_{x=0}^{n} \left(x(x-1)\binom{n}{x}p^x(1-p)^{n-x} + x\binom{n}{x}p^x(1-p)^{n-x}\right) \\
&= \sum_{x=0}^{n} \left(x(x-1)\binom{n}{x}p^x(1-p)^{n-x} + \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x}\right)
\end{aligned}$$

Notice the second summation in the last line is exactly $\mathbb{E}[X]$, so we already know this equals $np$. To compute the first summation we perform exactly the same sort of manipulation we did in the proof of

Proposition 9.2:

$$\begin{aligned}
&\sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x(1-p)^{n-x} \\
=&\sum_{x=0}^{n} x(x-1)\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} \\
=&\sum_{x=2}^{n} x(x-1)\frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} \\
=&\sum_{x=2}^{n} \frac{n(n-1)(n-2)!}{(n-2-(x-2))!(x-2)!}p^2p^{x-2}(1-p)^{n-2-(x-2)} \\
=&n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(n-2-(x-2))!(x-2)!}p^{x-2}(1-p)^{n-2-(x-2)} \\
=&n(n-1)p^2\sum_{x=0}^{n-2} \binom{n-2}{x}p^x(1-p)^{n-2-x} \\
=&n(n-1)p^2
\end{aligned}$$

Plugging this into the above we have

$$\mathbb{E}[X^2] = n(n-1)p^2 + np = n^2p^2 - np^2 + np.$$

Thus the variance is

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np - np^2 \\
&= np(1-p).
\end{aligned}$$

□

**Corollary 9.4.**
*If* $X \sim \mathrm{Binomial}(n,p)$*, then the standard deviation of* $X$ *is* $\sqrt{np(1-p)}$.

## 9.3 Geometric

Geometric random variables arise from the following type of experiment. Suppose we perform trials which are classified as success or failure, and all trials are independent of one another. In the case of a binomial random variable we perform some fixed number of trials, then count the successes. In a geometric random variable, however, we repeatedly perform the trials *until* we achieve a first success, and then count the number of trials which were required. For example, we may flip a coin repeatedly until it first lands on heads. If the first heads occurred on the fifth flip, then the value of our random variable would be five. Of course, there is a parameter hiding in this description of the geometric random variable: the probability of success on each trial, which we denote by $p$.

So, suppose $X$ was such a random variable. What should the pmf of $X$ be? If we get our first success on the $x$-th trial, then our previous $x-1$ trials must have been failures. Each of these failures has probability $1-p$, and the probability of success on that $x$-th trial is $p$, so the pmf is

$$p_X(x) = \begin{cases} (1-p)^{x-1}p & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases}$$

A random variable $X$ with such a pmf is called a ***geometric random variable with parameter*** $p$, and we write $X \sim \text{Geom}(p)$.

**Example 9.3.**
When rolling a six-sided die, the probability of getting a particular number, say 5, is $1/6$. What is the probability if you roll the die repeatedly the first time you roll a 5 is on the ninth roll?

Here we have a geometric random variable with parameter $p = 1/6$ which counts the number of required rolls to obtain a 5; $X \sim \text{Geom}(1/6)$. The probability the five first appears on the ninth roll is

$$p_X(9) = (1-1/6)^8 \cdot 1/6 = \frac{5^8}{6^9} = \frac{390625}{10077696} \approx 0.0388.$$

Of course, we want to know what the expected value of $X \sim \text{Geom}(p)$ is as a function of $p$. To do this we will need to evaluate a geometric series (and this is why this is called a geometric random variable). Recall that

if $a$ is any real number and $r$ is a real number satisfying $|r| < 1$, then the geometric series

$$\sum_{n=0}^{\infty} ar^n$$

converges to $\frac{a}{1-r}$. You should have seen this in a second semester calculus class, but just for the sake of completeness we'll provide a proof.

**Lemma 9.5.**
*If $a$ is any real number and $r$ is a real number satisfying $|r| < 1$, then*

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}.$$

*Proof.*
We first note that since there is an $a$ in each term of the series,

$$\sum_{n=0}^{\infty} ar^n = a + ar + ar^2 + ar^3 + \cdots$$

we can factor the $a$ out:

$$\sum_{n=0}^{\infty} ar^n = a\sum_{n=0}^{\infty} r^n = a(1 + r + r^2 + r^3 + \cdots).$$

So it suffices to see that $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$, then we can just multiply through by $a$. To see this series converges to claimed sum, recall that a convergent series equals the limit of partial sums,

$$\sum_{n=0}^{\infty} r^n = \lim_{N\to\infty} \sum_{n=0}^{N} r^n = \lim_{n\to\infty} \left(1 + r + r^2 + \cdots + r^N\right).$$

To determine this limit, let $S_N$ denote the $N$-th partial sum,

$$S_N = 1 + r + r^2 + \cdots + r^N.$$

If we multiply both sides by $r$ have

$$rS_N = r + r^2 + r^3 + \cdots + r^{N+1}.$$

Now consider the difference between $S_N$ and $rS_N$,

$$S_N - rS_N = (1 + r + \cdots + r^N) - (r + r^2 + r^{N+1}) = 1 - r^{N+1}.$$

Factoring $S_N$ from the left and dividing we have

$$\begin{aligned} S_N - rS_N &= 1 - r^{N+1} \\ \implies S_N(1 - r) &= 1 - r^{N+1} \\ \implies S_N &= \frac{1 - r^{N+1}}{1 - r}. \end{aligned}$$

That is,

$$\sum_{n=0}^{N} r^n = \frac{1 - r^{N+1}}{1 - r}.$$

Now we take the limit as $N$ goes to infinity. Obviously, since only one term in the expression above depends on $N$ we only need to compute $\lim_{N\to\infty} r^{N+1}$. Since $|r| < 1$, however, $r^{N+1}$ goes to zero as $N$ goes to infinity. □

**Proposition 9.6.**
*If $X \sim \text{Geom}(p)$, then $\mathbb{E}[X] = 1/p$.*

*Proof.*
We begin by writing out the definition of the expected value and factoring out the $p$ that appears in each term,

$$\mathbb{E}[X] = \sum_{x\in\mathbb{R}} x p_X(x) = \sum_{x=1}^{\infty} x(1-p)^{x-1}p = p\sum_{x=1}^{\infty} x(1-p)^{x-1}$$

Now we do something a little unexpected. Recall that the derivative of $t^n$, with respect to $t$, is $\frac{d}{dt}t^n = nt^{n-1}$. Differentiating $(1-t)^n$ likewise results in $\frac{d}{dt}(1-t)^n = -n(1-t)^{n-1}$ because of the chain rule. Notice this is very similar to the $x(1-p)^{x-1}$ we are summing above, except that we're missing a negative sign, but that's easy to compensate for.

What we'll do, then, is think of the terms $x(1-p)^{x-1}$ as being the derivative of $-(1-p)^x$ with respect to $p$. (Notice we're treating $p$ as the variable, not $x$, so we don't have to do logarithmic differentiation to compute the above derivative.) Thus

$$\begin{aligned}
\mathbb{E}[X] &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \\
&= p\sum_{x=1}^{\infty} -\frac{d}{dp}(1-p)^x \\
&= -p\frac{d}{dp}\sum_{x=1}^{\infty}(1-p)^x \\
&= -p\frac{d}{dp}\left(\sum_{x=0}^{\infty}(1-p)^x - 1\right) \\
&= -p\frac{d}{dp}\left(\frac{1}{1-(1-p)} - 1\right) \\
&= -p\frac{d}{dp}\left(\frac{1}{p} - 1\right) \\
&= -p\frac{d}{dp}\left(p^{-1} - 1\right) \\
&= -p\cdot\left(-p^{-2}\right) \\
&= \frac{p}{p^2} \\
&= \frac{1}{p}
\end{aligned}$$

□

**Exercise 9.2.**
Show that geometric random variables have the following "memory-

less" property: if $X \sim \text{Geo}(q)$, then for any integers $m > n > 0$,

$$P(X > m | X > n) = P(X > m - n).$$

## 9.4 Hypergeometric

A hypergeometric random variable, like a binomial random variable, counts the number of "successes" observed from a collection of $n$ trials. Unlike the binomial, however, the probability of successes changes from trial to trial.

Imagine, for example, a collection of 500 students is given and 30 of these students are math majors. If we choose ten students at random, we may want to know how many of the selected students are math majors (these are the "successes" in our trials). In such an example we have a ***population*** (the 500 students), from which we select a ***sample*** (the 10 randomly chosen students), and we want to know the number of selections which satisfied some criterion (being a math major). What makes this different from the binomial is that the probabilities change: the first student we select has chance $30/500$ of being a math major, but what about the second student? If our first student was a math major, the chance the second one is as well is $29/499$, but if the first student was not a math major, then the probability will be $30/499$.

Let's reason our way through the probability there will be exactly one math major among the ten students we pick. Of the thirty math majors we have to choose one, and there are $\binom{30}{1}$ ways to do this. Now we need to choose the remaining students. Since we want to only have one math major, the other nine students must be non-math majors. Since 30 of our 500 students were math majors, 470 are non-math majors and we need to choose nine of them. Of course, there $\binom{470}{9}$ ways to do this. So, the number of ways we can choose ten students with exactly one math major is $\binom{30}{1}\binom{470}{9}$. The total number of ways we can choose ten students from the 500 students is $\binom{500}{10}$. Hence the probability of exactly one math major when we choose ten students is

$$\frac{\binom{30}{1}\binom{470}{1}}{\binom{500}{10}}.$$

Similarly, the probability we choose exactly two math majors when we choosen ten students from a population of 500 students containing a total

of 30 math majors is

$$\frac{\binom{30}{2}\binom{470}{8}}{\binom{500}{10}}.$$

A general hypergeometric random variable simply generalizes this situation. Note we have a few parameters here:

- The population size, say $N$.
- The sample size, say $n$.
- The number of "successes" in the population, say $k$.

To denote that $X$ is a random variable counting the number of successes in a sample of size $n$ coming from a population of size $N$ which contains $k$ total successes, we write $X \simeq \text{Hyp}(N, n, k)$ and call $X$ a ***hypergeometric random variable***. Generalizing the argument above about picking math majors, we see that the pmf of such a random is

$$p_X(x) = \begin{cases} \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} & \text{if } x \in \mathbb{Z} \cap [\max(0, n+k-N), \min(n,k)] \\ 0 & \text{otherwise} \end{cases}$$

Most of the formula for the pmf above should make sense, except possible the conditions given for the first part of the pmf, the $\mathbb{Z} \cap [\max(0, n+k-N), \min(n,k)]$. What we're saying here is that in order for $p_X(x)$ to have any hope of not being zero, we need that $x$ is an integer which is no smaller than $\max(0, n+k-N)$, and no larger than $\min(n,k)$. Since we're counting the number of successes, of course $x$ needs to be an integer (e.g., we can't have $\sqrt{2}$ successes). The number of successes can't exceed the number of samples we take (i.e., $x$ can be no larger than $n$), nor can it exceed the total number of successes in the population ($x$ can not be larger than $k$); together these mean $x$ can be no larger than the minimum of $n$ and $k$.

To understand the $\max(0, n+k-N)$ part, notice that if our sample size is large enough we may be forced to have at least a few successes. For example, if we had a population of size 10 and 8 elements of the population were successes, then any sample of size 3 or more must have at least one success. So, what's the fewest number of successes we can possible have in general, when the population has size $N$, the sample has size $n$, and the number of successes in the population is $k$? Notice that if the number of failures (non-successes) in the population is smaller than our sample size, we must have some successes in any sample. The number of failures is $N-k$, so if we have a sample of size $n > N-k$ we must have at least $n-(N-k)$

successes, which can rewrite as $n + k - N$. Of course, when $n \leq N - k$ this quantity is negative which doesn't make sense when we're counting the number of successes. If we want $x$ to be no smaller than zero and no smaller than $n + k - N$, a succinct way to write this is $x \geq \max(0, n + k - N)$.

**Example 9.4.**
In *Texas Hold 'Em*, each player receives two cards and then three cards are turned face up on the table. If the player received two 2's, and the face-up cards are 3, $J$, and $A$, what is the probability we will be able to make two pair with our two 2's and one of the face-up cards when two more face-up cards are added to the three currently on the table?

Here the population we are interested in is the remaining 47 cards (52 cards to start, minus the 5 that have been dealt). Our sample size is 2 since two more cards will be revealed, and the number of successes (the cards which will allow us to make two pair) is 9 since there are three remaining 3's, three $J$'s, and three $A$'s. That is, we have a hypergeometric random variable with $N = 47$, $n = 2$, and $k = 9$. Hence the probability one of the two cards will give us a two-pair is

$$\frac{\binom{9}{1}\binom{38}{1}}{\binom{47}{2}} = \frac{342}{1041} \approx 0.3285.$$

(Here we're considering the situation where exactly one of the two remaining revealed cards allows us to build a two pair. To make the calculation easier we are ignoring a few situations, such as if two cards of the same rank, other than the ranks already revealed, appeared. Adding those cases into our calculation, the probability of making a two pair is actually higher than the number calculated above. This is why getting two cards of the same rank is a very nice hand in Texas Hold 'Em.)

## 9.5 Negative binomial

In the case of a geometric random variable we said the underlying experiment was that trials are repeated until a success is obtained. In the negative binomial something similar happens: we again repeat independent trials

which are success with probability $p$ and failture with probability $1-p$, but we repeat the trials until some fixed number of successes, say $r$, are obtained, and count the the number of failures that occurred before obtaining the $r$-th success.

(In the binomial we fix the number of trials and count the number of successes, whereas in the negative binomial we fix the number of successes and count the number of failures before seeing the prescribed number of successes.)

Suppose it took $x$ failures before obtaining $r$ successes. Notice that the very last trial is a success since we stop the experiment once we have $r$ successes. So all we need to do is determine which $x$ of the earlier $x+r-1$ trials were failures, then multiply the probabilities of that number of successes and failures. That is, the probability of $x$ failures before $r$ successes is

$$\binom{x+r-1}{r-1} p^r (1-p)^x$$

In such a situation we say that $X$ is a ***negative binomial*** random variable with parameters $p$ and $r$ and write $X \sim \text{NegBin}(p, r)$, and the pmf of such a random variable is

$$p_X(x) = \begin{cases} \binom{x+r-1}{r-1} p^r (1-p)^x & \text{if } x \in \mathbb{N} \cup \{0\} \\ 0 & \text{otherwise} \end{cases}$$

**Example 9.5.**
Suppose a paleontologist wants to collect fossils of dinosaurs until they find three fossils with evidence of dinosaur feathers. If only $1/10$ of of fossils have evidence of feathers, what is the probability the paleontologist finds 15 fossils without evidence of feathers before finding three fossils with evidence?

Here we have a negative binomial with parameters $p = 1/10$ and $r = 3$. By the above we know

$$\begin{aligned} p_X(15) &= \binom{15+3-1}{15} \left(\frac{1}{10}\right)^3 \left(\frac{9}{10}\right)^{15} \\ &\approx 0.028 \end{aligned}$$

## 9.6 Poisson

Imagine that you are interested in counting the number of times some random phenomenon occurs over a given length of time. For example, maybe you count the number of fish that swim under a bridge over the course of ten minutes; or you count the number of students that walk by the Sample Gates over an hour; or you count the number of traffic accidents that occur in a city over the course of a month.

Now suppose that the events you're counting are independent of one another. That is, one car accident on one day doesn't imply anything about whether there will be any more car accidents that day.

A reasonable way to try to compute the probability of $k$ of these random occurences in a length of time as follows. Say that from previous observations you know the average number of occurences in the given length of time. For example, maybe from past experience you know on average fifteen fish swim under the bridge evern ten minutes; or 230 students walk by the Sample Gates every hour on average; or there are an average of twenty two traffic accidents per month. Call this average value $\lambda$.

Now suppose that you break your time interval up into $n$ discrete chunks (e.g., seconds, minutes, or days), and you mark each "chunk" of time as either a success or failure depending on whether the random occurence we're interested in occurred or not. For example, if we are interested in the number of fishing swimming under a bridge over the course of ten minutes, we might break our ten minutes up into ten one-minute chunks and count each chunk as being a success if a fish swam by during that minute, and a failure otherwise.

By marking each chunk of time as success or failure, we turn our original experiment into a binomial random variable. Our $n$ chunks of time become $n$ trials marked as success or failure. Recall that each trial in a binomial has equal probability $p$ of being a success. What should $p$ be in our situation? Since a binomial with parameters $n$ and $p$ has expected value $np$ and since we know the average number of occurences is $\lambda$, we must have $\lambda = np$ and so $p = \lambda/n$.

Now the probability of $k$ of our $n$ chunks of time being marked as successes is

$$\binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Notice there's a bit of a problem with our counting the number of random occurrences using binomial random variables in the way outlined above: when we mark each chunk of time as success or failure, we don't have any way of distinguishing one success during that chunk of time from

two successes or three successes or four successes ... That is, our binomial random variable is really more of a very rough estimate for counting the number of occurrences. We can improve our estimate, though, by using more (and hence smaller) chunks of time.

For example, when counting fish swimming under the bridge, if we break our ten minute interval up into ten one minute chunks and three fish swim under the bridge during one of those minutes, just marking the chunk as success or failure doesn't include this information. If we break into smaller chunks, say 600 seconds, then our three fish swimming by are more likely to be distinguished as three different successes (as long as they don't swim by during the same second) than before. Of course it could happen that multiple fish swim by in one second, so we get an even better estimate by using even smaller intervals.

Ultimately what we want to do is have infinitesimally small units of time by taking the limit as the number of chunks of time goes to infinity. We can actually do this, but we'll need to do some algebraic manipulations and recall a basic fact from calculus in order to actually make all of this precise.

First let's take our expression for the probability of $k$ successes when we divide our unit of time into $n$ equal pieces,

$$\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

and rewrite it follows just by writing out $\binom{n}{k}$ in terms of factorials and rewriting the other two factors raised to powers with basic properties from algebra:

$$\frac{n!}{(n-k)!k!}\cdot\frac{\lambda^k}{n^k}\cdot\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k}.$$

Now let's cancel out terms in the $\frac{n!}{(n-k)!}$ above and also swap the $n^k$ and $k!$ in the denominators of the first two factors to obtain

$$\frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\cdot\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k}.$$

Now we want to take the limit of this as $n$ goes to infinity. We'll do this factor by factor.

For the first factor, note that if we multiply out

$$n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1)$$

then we'd have an expression of the form

$$n^k+(\text{stuff of degree less than } k).$$

This is simply because there are $k$ factors in the expression above and they all have the form ($n$ − something). We could work out exactly what all of the "stuff" alluded to above is, but it won't matter because in the limit we will have

$$\lim_{n\to\infty} \frac{n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1)}{n^k} = \lim_{n\to\infty} \frac{n^k + (\text{stuff of degree less than } k)}{n^k}.$$

As $n$ goes to infinity, the $n^k$ term in the numerator grows so much faster than the other terms that the other terms become irrelevant. (If you want to be precise, multiply and divide by $1/n^k$ or apply l'Hôpital's rule $k$ times.) That is, as $n$ gets really big, the fraction essentially becomes $\frac{n^k}{n^k}$, and so in the limit this is just 1.

The next factor $\frac{\lambda^k}{k!}$ has no $n$'s in it, so nothing happens to this as $n$ goes to infinity.

For the last factor we take the limit of the denominator and numerator separately. For the denominator notice that

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^k = 1$$

since $\lambda/n \to 0$ as $n \to \infty$.

Finally for limit of the numerator of the last factor we need the following factoids from calculus.

**Lemma 9.7.**
*For any real number $x$,*

$$\lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n = e^x.$$

*Proof.*
First recall that the Taylor series for $e^x$ is

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

So we want to show that as $n$ gets very large, $(1+x/n)^n$ looks like this

Taylor series. To do this we recall that the binomial theorem states

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

Applying this to $(1+x/n)^n$ we have

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

Above we already discussed what happens to $\frac{n!}{n^k(n-k)!}$ as $n$ goes to infinity: this approaches 1. That is, for very, very large values of $n$ we have

$$\left(1+\frac{x}{n}\right)^n \approx \sum_{k=0}^{n} \frac{x^k}{k!}.$$

So as $n$ goes to infinity, $(1+x/n)^n$ goes to the Taylor series of $e^x$. □

With Lemma 9.7 at our disposal, we now easily see

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n = \lim_{n\to\infty} \left(1+\frac{-\lambda}{n}\right)^n = e^{-\lambda}.$$

Putting all of this together, we see that the probability of $k$ random occurences over a length of time where all occurences are independent and the average number of occurences in the time interval is $\lambda$ is

$$e^{-\lambda}\frac{\lambda^k}{k!}.$$

This gives the pmf of a family of random variables called the ***Poisson random variables***. We write $X \sim \text{Poisson}(\lambda)$ when $X$ has pmf

$$p_X(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & \text{if } x = 0,1,2,3,... \\ 0 & \text{otherwise} \end{cases}$$

**Example 9.6.**
Suppose the number of typos on a single page of a book is, on average, one typo per page. Modelling the number of typos by a Poisson random variable, what is the probability of exactly two typos? What is the probability of *at least* one typo on a given page?

The number of typos on a page is counted by the random variable $X \sim \text{Poisson}(1)$. The probability of exactly two typos is

$$\Pr(X = 2) = p_X(2) = e^{-2}\frac{1^2}{2!} = \frac{1}{2e} \approx 0.1839.$$

The complement of at least one typo is zero typos. The probability of zero typos is

$$\Pr(X = 0) = p_X(0) = e^{-1}\frac{1^0}{0!} = \frac{1}{e} \approx 0.3679,$$

so the probability of at least one typo is

$$\Pr(X \geq 1) = 1 - \Pr(X = 0) = 1 - \frac{1}{e} \approx 0.6321.$$

**Exercise 9.3.**
Suppose $X \sim \text{Poisson}(\lambda)$ for some $\lambda > 0$. Verify that $\mathbb{E}[X] = \lambda$.

**Example 9.7.**
Suppose the number of automobile accidents in a certain city is on average three per week. Modelling this as a Poisson random variable, what is the probability there are no accidents in a given week?

The number of accidents is counted by $X \sim \text{Poisson}(3)$, and so the probability of zero accidents is

$$\Pr(X = 0) = p_X(0) = e^{-3}\frac{3^0}{0!} = e^{-3} \approx 0.04978.$$

So there's just shy of a 5% chance there will be no accidents in a given week.

**Exercise 9.4.**
Suppose $X \sim \text{Poisson}(\lambda)$ for some $\lambda > 0$. Verify that $\text{Var}[X] = \lambda$.

When we use a Poisson random variable we make a choice about how big our unit of time should be, and the parameter $\lambda$ represents the average number of occurences in this length of time. If we want to use a different length of time, we need to scale $\lambda$ appropriately. For example, if as in Example 9.7, the number of accidents per week is Poisson(3), then the number of accidents per two weeks is given by Poisson(6), the number of accidents per day is Poisson($3/7$), and the number of accidents per year is Poisson(156).

In general, if the number of random occurences our Poisson random variable counts is on average $\lambda$ per unit time, then the number of random occurences in an interval of time of length $t$ is Poisson with parameter $\lambda t$.

**Example 9.8.**
What is the probability that in a city with an average of three automobile accidents per week there are zero accidents on any given day.

As noted above this is counted by $X \sim \text{Poisson}(3/7)$, and so the probability of zero accidents on any given day is

$$e^{-3/7}\frac{(3/7)^0}{0!} = e^{-3/7} \approx 0.6514.$$

Notice that when we have one Poisson random variable $X \sim \text{Poisson}(\lambda)$, there's a natural way to define an infinite family of Poisson random variables: for each $t > 0$ we define $X_t \sim \text{Poisson}(\lambda t)$.

**Remark.**
An infinite family of random variables like this is called a ***stochastic process***, and the study of stochastic processes is an interesting realm of modern mathematics that has applications to finance, physics, computer science, and other disciplines.

You may have seen another type of stochastic process in other probability classes before: a Markov chain represents the state of a system at a point in time, where the probability of the system's next state depends only on the current state. Letting $X_1$, $X_2$, $X_3$, ... denote the state of the system at time 1, time 2, time 3, ... gives a stochastic process.

## 9.7 Practice problems

**Problem 9.1.**
Suppose a multiple choice exam has five questions, and each question has three possible answers. If a student were to randomly guess the answers to each question (assuming each of the three possibilities is equally likely to be selected by the student), what is the probability the student would get at least four answers correct?

**Problem 9.2.**
Suppose that a factory produces piston heads for car engines. In order for the piston to work correctly, the head must be very close to circular: the engine will not work correctly if the piston head is not within a certain tolerance of being a perfect circle. Suppose eight piston heads are selected from a batch of fifty and it is known that three of the fifty heads are not circular enough to work in a particular type of engine. What is the probability that exactly six of the selected heads will work in the engine?

**Problem 9.3.**
A *Hamming code* is a type of error-correcting code often used in telecommunications to reduce the likelihood of receiving a corrupted message. For example, a four-bit message can be encoded using seven bits in such a way that if one of the bits is corrupted (e.g., a transmission error causes a 1 to flip to a 0 or vice versa), the original, intended message can still be reconstructed. If two or more bits are corrupted, however, then the entire message is corrupted.

If the probability any one bit is corrupted is $1/10$, what's the probability the original four-bit message can be reconstructed from a seven bit Hamming code?

# 10 Continuous Random Variables

*"Obvious" is the most dangerous word in mathematics.*

E. T. Bell

## 10.1 Introduction

When we discussed discrete random variables we saw that probabilities could be computed using a probability mass function, $p_X$, whose value $p_X(x)$ told us the probability $\Pr(X = x)$. There are some random variables where $\Pr(X = x)$ will be zero for every value of $x$, however. For example, imagine our underlying experiment is throwing a random dart at a circular dart board of radius 1. Now associate a random variable $X$ to this experiment where $X$ associates to each point on the board the distance from that point to the origin. To find the probability $\Pr(X = 1/2)$, we would need to find the probability a dart lands distance exactly $1/2$ from the origin. That is, the dart would need to land on the circle of radius $1/2$. What's the probability this happens? We compute the probability of landing in a region on the board by dividing the area of that region by the area of the entire board. The circle, however, has zero area, and so the probability is zero. Of course, there's nothing magical about $1/2$ in this example: for any value of $x$, $\Pr(X = x)$ will be zero. That is, for a random variable such as this the idea of a probability mass function isn't very helpful: the function would just be constant zero everywhere!

If we can't make sense of a probability mass function here, how should we try to compute probabilities? Recall that we had another function we could associate to discrete random variables: the cumulative distribution function $F$ was defined by $F(x) = \Pr(X \leq x)$. Would this function be helpful in our dart board example? Notice that $F(1/2)$ is $\Pr(X \leq 1/2)$ which would be the area of the disc of radius $1/2$ divided the area of the whole board:

$$F(1/2) = \Pr(X \leq 1/2) = \frac{\pi/4}{\pi} = \frac{1}{4}.$$

So, even though a probability mass function doesn't make sense for our random variable above, the cumulative distribution function does.

Now let's make an observation about this cumulative distribution function. Notice for any value of $x$ between 0 and 1, the cdf for our random

variable above is

$$F(x) = \Pr(X \leq x) = \frac{\pi x^2}{\pi} = x^2.$$

Additionally, $F(x) = 1$ for any $x \geq 1$ (since the dart board only has radius 1) and $F(x) = 0$ for any $x \leq 0$ (since our distance to the origin of the board will never be negative). So the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^2 & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Notice this is a continuous function (it's not differentiable at $x = 0$ or $x = 1$, but it's still continuous). If a random variable $X$ has a continuous cdf, then we call $X$ a ***continuous random variable***.

**Remark.**
Notice that discrete random variables are never continuous (using this definition of continuous random variable): the cdf of a discrete random variable has a jump discontinuity at every value of $x$ for which the pmf $p_X(x)$ is non-zero.

Let's notice too that even though our cdf $F(x)$ above is not differentiable everywhere, it's differentiable everywhere except at two distinct points. So, we can define its derivative, which we'll call $f(x)$, at every point except $x = 0$ and $x = 1$. This gives us

$$f(x) = F'(x) = \begin{cases} 0 & \text{if } x < 0 \\ 2x & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$$

Notice that, because of the fundamental theorem of calculus, we can calculate probabilities by integrating $f$:

$$\Pr(a \leq X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

Even though $f$ is not defined at every point, this integral still makes sense: you could fill in the "holes" and define $f(0)$ and $f(1)$ to be anything you'd

like and it wouldn't affect the value of the integral. (This is not true for arbitrary functions because we could have vertical asymptotes and might worry about having an improper integral. However, if a function is are the derivative of the cdf of a random variable, this won't happen because the $0 \leq F(x) \leq 1$ for every $x$.)

## 10.2 Probability density and cumulative distribution

A function, such as our $f$ in the discussion above, which has the property that

$$Pr(a \leq X \leq b) = \int_a^b f(x)\,dx$$

is called a ***probability density function (pdf)*** for the random variable $X$.

Notice the pdf $f$ of a continuous random variable $X$ must have the following two properties:

1. $f(x) \geq 0$ for all $x$, and
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

These properties follow simply from the fact that probabilities are never negative, and $\Pr(\Omega) = 1$ for any sample space $\Omega$ and probability function Pr.

Just as we can specify a discrete random variable by giving its pmf, we can specify a continuous random variable by giving its pdf. That is, any integrable function $f(x)$ satisfying the two properties above is the pdf of some random variable $X$. Said another way, if we know the pdf of $X$ then we know everything we need to know about $X$; we don't need to know anything about the underlying experiment or sample space, we only need the pdf.

**Example 10.1.**
Verify that the function

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

is the pdf of some random variable $X$, then compute $\Pr(1 \le X \le 3/2)$.

First we must verify the two conditions above. Obviously $f(x) \ge 0$ for all $x$, and integrating the function gives

$$\begin{aligned}
&\int_{-\infty}^{\infty} f(x)\,dx \\
=&\int_{-\infty}^{0} 0\,dx + \int_{0}^{2} \frac{3x^2}{8}\,dx + \int_{2}^{\infty} 0\,dx \\
=&\frac{x^3}{8}\Big|_0^2 \\
=&1.
\end{aligned}$$

Now to compute $\Pr(1 \le X \le 3/2$, we simply integrate $f(x)$:

$$\begin{aligned}
\Pr(1 \le X \le 3/2) =& \int_{1}^{3/2} f(x)\,dx \\
=& \int_{1}^{3/2} \frac{3x^2}{8}\,dx \\
=& \frac{x^3}{8}\Big|_1^{3/2} \\
=& \frac{27}{64} - \frac{1}{8} \\
=& \frac{27-8}{64} \\
=& \frac{19}{64} \approx 0.2969
\end{aligned}$$

**Example 10.2.**
What choice of constant $k$ makes the function

$$f(x) = \begin{cases} k[1-(x-3)^2] & \text{if } 2 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

a pdf?

Notice that $k[1-(x-3)^2]$ is non-negative for $x \in [2,4]$, so as long as $k \geq 0$, $f(x)$ will be non-negative everywhere. We need to find the choice of $k$ that makes the function $f(x)$ above integrate to 1. First we compute this integral as a function of $k$:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{2}^{4} k[1-(x-3)^2]\,dx.$$

Performing the substitution $u = x-3$, $du = dx$ the integral becomes

$$k\int_{-1}^{1}(1-u^2)\,du = k\left(u - \frac{u^3}{3}\right)\Big|_{-1}^{1} = k\left(1-\frac{1}{3}\right) - k\left(1-\frac{-1}{3}\right) = \frac{4k}{3}.$$

Setting this equal to 1 and solving for $k$ gives $k = \frac{3}{4}$.

The ***cumulative distribution function (cdf)*** of a continuous random variable, just as for a discrete random variable, is defined to be the function $F(x)$ determined by

$$F(x) = \Pr(X \leq x).$$

In terms of the pdf $f(x)$, the cdf can be calculated as

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

By the fundamental theorem of calculus, this means the cdf is an antiderivative of the pdf.

In the case of our random variable $X$ with pdf

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^3}{8} & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2. \end{cases}$$

**Example 10.3.**
What is the cdf of the continuous random variable $X$ with pdf

$$f(x) = \begin{cases} 0 & \text{if } x < -2 \\ \frac{5x^4}{64} & \text{if } -2 \leq x \leq 2 \\ 0 & \text{if } x > 2 \end{cases}$$

We simply need to integrate $f(x)$ to find the cdf $F(x)$. Notice that since the pdf is a piecewise function defined on three particular intervals, $(-\infty, -2)$, $[-2, 2]$, and $(2, \infty)$, we should expect the cdf to likewise be a piecewise function defined on these three intervals.

For $x \in (-\infty, -2)$, we have

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{x} 0\,dt = 0.$$

For $x \in [-2, 2]$ we have

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-\infty}^{-2} f(t)\,dt + \int_{-2}^{x} f(t)\,dt \\ &= \int_{-\infty}^{-2} 0\,dt + \int_{-2}^{x} \frac{5t^4}{64}\,dt \\ &= \frac{t^5}{64}\Big|_{-2}^{x} \\ &= \frac{x^5}{64} - \frac{(-2)^5}{64} \\ &= \frac{x^5 + 32}{64}. \end{aligned}$$

Finally, for $x > 2$ we have

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,dt \\ &= \int_{-\infty}^{-2} f(t)\,dt + \int_{-2}^{2} f(t)\,dt + \int_{2}^{x} f(t)\,dt \\ &= 0 + 1 + \int_{2}^{x} 0\,dt \\ &= 1 \end{aligned}$$

Putting all of this together, the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < -2 \\ \frac{x^5+32}{64} & \text{if } -2 \leq x \leq 2 \\ 1 & \text{if } x > 2 \end{cases}$$

## 10.3 Percentiles

Since the cdf of a continuous random variable is continuous and has the properties that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$, the intermediate value theorem tells us that for every value of $p$ in the interval $(0, 1)$, there must exist a value of $\eta$ such that $F(\eta) = p$. Of course, $\eta$ depends on $p$ so we should this as $\eta(p)$. This value of $\eta$ for a given choice of $p$ is called a percentile.

More precisely, for each $p \in (0, 1)$ we define the $(100 \cdot p)$***-th percentile*** of a continuous random variable $X$ to be the real number $\eta(p)$ such that $F(\eta(p)) = p$. For example, when $p = 0.5$, the 50-th percentile is the value of $\eta(p)$ such that $F(\eta(p)) = 0.5$; when $p = 0.3$ the 30-th percentile is the value of $\eta(p)$ such that $F(\eta(p)) = 0.3$. The 50-th percentile of a continuous random variable is often called the ***median***.

**Remark.**
Notice that the notion of percentile doesn't really make sense for discrete random variables, at least not for all choices of $p$, since the cdf can "jump over" values of $p$. For instance, suppose $X$ is a discrete

random variable with pmf $p(x)$

$$p(x) = \begin{cases} 0.3 & \text{if } x = 1 \\ 0.7 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

The cdf is then

$$F(x) = \begin{cases} 0 & \text{if } x < 1 \\ 0.3 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2. \end{cases}$$

Here there is no value of $\eta$ such that $F(\eta) = 0.5$, for example, and so there is no notion of a 50-th percentile.

**Example 10.4.**
Consider the continuous random variable $X$ with pdf

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

What is the 90-th percentile of $X$?

Since percentiles are defined in terms of the cdf, we first need to compute the cdf of this random varible. We had already done this above, however, and found the cdf was

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^3}{8} & \text{if } 0 \leq x \leq 2 \\ 1 & \text{if } x > 2. \end{cases}$$

We need to find the value of $\eta$ such that $F(\eta) = 0.9$. Since $F(0) = 0$ and $F(2) = 1$, it's clear that this $\eta$ must occur between 0 and 2, and

so $F(\eta) = \frac{\eta^3}{8}$. Now this is a simple algebra problem:

$$\begin{aligned} & F(\eta) = 0.9 \\ \implies & \frac{\eta^3}{8} = 0.9 \\ \implies & \eta^3 = 0.9 \cdot 8 = 7.2 \\ \implies & \eta = \sqrt[3]{7.2} \approx 1.931. \end{aligned}$$

So the 90-th percentile is $\sqrt[3]{7.2} \approx 1.931$.

## 10.4 Expected value

Recall that for a discrete random variable $X$ we defined the expected value of $X$ as a weighted average of the values $X$ could take on times the probability $X$ takes on those values: $\mathbb{E}[X] = \sum_{x \in \mathbb{R}} xp(x)$. In the case of discrete random variables the sum above is well-defined since $p(x)$ equals zero for "most" values of $x$ and the sum becomes either a simple sum of finitely-many terms or an infinite series.

For continuous random variables we would like to define something similar, but there are two issues. First, we don't have a probability mass function. The density function for a continuous random variable is similar to the mass function for a discrete random variable, however, so this issue shouldn't bother us too much. The more serious issue is that the density function could be non-zero for uncountably many values of $x$, and so it's not really clear what a simple summation of $xf(x)$ taken over all values of $x$ should be.

Since we know how to compute expected values for discrete random variables and are trying to figure out what expected values are for continuous random variables, maybe we should try to approximate a continuous random variable by a discrete random variable, and then compute the expected value of that discrete random variable. We may then try to update improve our approximation and compute the expected value of this new, improved approximation. If we keep doing this over and over, do our expected values of the approximations converge to any one number? If so, maybe we should define the expected value of our continuous random variable to be that value the approximations converge to.

To get started, let's go back to our running example of the continuous random variable $X$ with pdf

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

To approximate this continuous random variable with a discrete random variable, maybe what we should do is chop the interval $[0, 2]$, where the density is non-zero, up into finitely-many subintervals, pick one point in each subinterval, and define the probability of this point as the probability $X$ takes on a value in that subinterval.

For example, maybe we chop $[0, 2]$ up into four subintervals all of width $1/2$ and pick one point in each interval, say the left-hand endpoint of that interval. That is, we'll build a discrete random variable whose probability mass function is non-zero at $x = 0$, $x = 1/2$, $x = 1$, and $x = 3/2$. We need to assign probabilities to these values, though, so let's define them to be the probability the original random variable took on a value in the interval $[0, 1/2)$, $[1/2, 1)$, $[1, 3/2)$, and $[3/2, 2]$. We of course compute these values by integrating the density function over these intervals:

$$\begin{aligned} \Pr(0 \leq X < 1/2) &= \int_0^{1/2} \frac{3x^2}{8}\, dx = 1/64 \\ \Pr(1/2 \leq X < 1) &= \int_{1/2}^{1} \frac{3x^2}{8}\, dx = 7/64 \\ \Pr(1 \leq X < 3/2) &= \int_{1}^{3/2} \frac{3x^2}{8}\, dx = 19/64 \\ \Pr(3/2 \leq X < 2) &= \int_{3/2}^{2} \frac{3x^2}{8}\, dx = 37/64 \end{aligned}$$

This random variable has probability mass function

$$p(x) = \begin{cases} 1/64 & \text{if } x = 0 \\ 7/64 & \text{if } x = 1/2 \\ 19/64 & \text{if } x = 1 \\ 37/64 & \text{if } x = 3/2 \end{cases}$$

and the expected value is

$$0 \cdot 1/64 + 1/2 \cdot 7/64 + 1 \cdot 19/64 + 3/2 \cdot 37/64 = 39/32 = 1.21875$$

Let's momentarily call this discrete random variable we've construct $X_4$ since it was built from 4 subintervals where the density of $X$ was non-zero.

We could now construct a random variable $X_8$ using the same procedure, but using eight different intervals of equal width and this would give us a discrete random variable with pmf

$$\begin{cases} {}^{1}\!/_{512} & \text{if } x = {}^{0}\!/_{4} \\ {}^{7}\!/_{512} & \text{if } x = {}^{1}\!/_{4} \\ {}^{19}\!/_{512} & \text{if } x = {}^{2}\!/_{4} \\ {}^{37}\!/_{512} & \text{if } x = {}^{3}\!/_{4} \\ {}^{61}\!/_{512} & \text{if } x = {}^{4}\!/_{4} \\ {}^{91}\!/_{512} & \text{if } x = {}^{5}\!/_{4} \\ {}^{127}\!/_{512} & \text{if } x = {}^{6}\!/_{4} \\ {}^{169}\!/_{512} & \text{if } x = {}^{7}\!/_{4} \end{cases}$$

The expected value of this random variable is

$$\mathbb{E}[X_8] = \frac{175}{128} = 1.3671875.$$

Our goal is keep producing these discrete random variables, whose expected values we know how to compute, which approximate our continuous random variable, and see if the expected values converge.

Let's suppose we kept doing this process forever, letting $X_n$ denote the discrete random variables constructed as above but using $n$ subintervals of equal width. Let's let $x_i$ denote the left-hand interval of the $i$-th interval. The expected value of $X_n$ can then be written as

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} x_i \Pr(x_i \leq X < x_{i+1}) = \sum_{i=1}^{n} x_i \int_{x_i}^{x_{i+1}} f(t)\, dt.$$

Recall the mean value theorem for integrals says that if $f(t)$ is continuous on an interval $[a, b]$, then there exists a value of $c$ such that $f(c) = \frac{1}{b-a}\int_a^b f(t)\, dt$. Letting $c_i$ be the corresponding value for the integral $\int_{x_i}^{x_{i+1}} f(t)\, dt$ above we can write the expected value as

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} x_i f(c_i)(x_{i+1} - x_i).$$

Now we write $\Delta x_i = x_{i+1} - x_i$ to write

$$\mathbb{E}[X_n] = \sum_{i=1}^{n} x_i f(c_i)\Delta x_i.$$

Notice that if $f$ is continuous then $x_i \approx f(c_i)$ when the intervals are very small and so

$$\mathbb{E}[X_n] \approx \sum_{i=1}^{n} x_i f(x_i) \Delta x_i.$$

Finally, taking the limit as $n$ goes to infinity we see

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \lim_{n\to\infty} \sum_{i=1}^{n} x_i f(x_i) \Delta x_i = \int_0^2 x f(x)\, dx.$$

The limits of integration are 0 to 2 because we were breaking the interval 0 to 2 up into $n$ subintervals to construct our random variables above. Since $f(x)$ was zero outside of $[0, 2]$, we see this is equal to

$$\int_{-\infty}^{\infty} x f(x)\, dx$$

and in general for a continuous random variable $X$ with probability density function $f(x)$, we define the ***expected value*** of $X$ as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx.$$

**Example 10.5.**
For the continuous random variable $X$ with density

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

the expected value is

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,dx \\
&= \int_{-\infty}^{0} x f(x)\,dx + \int_{0}^{2} x f(x)\,dx + \int_{2}^{\infty} x f(x)\,dx \\
&= 0 + \int_{0}^{2} x\frac{3x^2}{8}\,dx + 0 \\
&= \frac{3x^4}{32}\Big|_0^2 \\
&= \frac{3\cdot 2^4}{32} - 0 \\
&= \frac{3}{2} = 1.5
\end{aligned}$$

**Example 10.6.**
The expected value of the continuous random variable $X$ with pdf

$$f(x) = \begin{cases} 0 & \text{if } x < -2 \\ \frac{5x^4}{64} & \text{if } -2 \le x \le 2 \\ 0 & \text{if } x > 2 \end{cases}$$

is

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,dx \\
&= \int_{-\infty}^{-2} x f(x)\,dx + \int_{-2}^{2} x f(x)\,dx + \int_{2}^{\infty} x f(x)\,dx \\
&= 0 + \int_{-2}^{2} x \cdot \frac{5x^4}{64}\,dx + 0 \\
&= \frac{5x^6}{384}\Big|_{-2}^{2} \\
&= \frac{5 \cdot 2^6 - 5 \cdot (-2)^6}{384} \\
&= \frac{320 - 320}{384} \\
&= 0.
\end{aligned}$$

The expected value of a continuous random variable has a very physical interpretation. Suppose the pdf $f(x)$ of a continuous random variable $X$ is non-zero only on inside interval $[a, b]$. Interpreting $f(x)$ as the density (in the sense of mass divided by length) of a rod of length $b - a$, $\mathbb{E}[X]$ is the center of mass of the rod: it is the point on the rod where a fulcrum could be placed and the rod would be perfectly balanced. Notice, however, this is usually *not* the same as the median of $X$ (aka the 50-th percentile), as the next example illustrates.

**Example 10.7.**
Consider the random variable $X$ with pdf

$$f(x) = \begin{cases} 9/10 & \text{if } 0 \le x \le 1 \\ 1/10 & \text{if } 10 \le x \le 11 \\ 0 & \text{otherwise} \end{cases}$$

It is easy to check that $5/9$ is the median:

$$\begin{aligned}\int_{-\infty}^{5/9} f(x)\,dx &= \int_{0}^{5/9} \frac{9}{10}\,dx \\ &= \frac{9x}{10}\Big|_0^{5/9} \\ &= \frac{9}{10}\cdot\frac{5}{9} \\ &= \frac{1}{2}.\end{aligned}$$

However, $\frac{5}{9}$ is not the expected value:

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,dx \\ &= \int_0^1 x\frac{9}{10}\,dx + \int_{10}^{11} x\frac{1}{10}\,dx \\ &= \frac{9x^2}{20}\Big|_0^1 + \frac{x^2}{20}\Big|_{10}^{11} \\ &= \frac{9}{20} + \frac{21}{20} \\ &= \frac{3}{2}\end{aligned}$$

## 10.5 Functions of random variables

Just as in the case of discrete random variables, we can compose a continuous random variable $X : \Omega \to \mathbb{R}$ with a function $g : \mathbb{R} \to \mathbb{R}$ to obtain a new random variable denoted $g \circ X$ or $g(X)$. With discrete random variables such a composition always produces a discrete random variable, but this not necessarily the case for a continuous random variable! That is, even if $X$ is a continuous random variable, it could be the composition $g(X)$ is discrete!

**Example 10.8.**

Suppose $X$ is the continuous random variable with pdf

$$f(x) = \begin{cases} \frac{x+1}{2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and let $g$ be the function

$$g(x) = \begin{cases} -1 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

Notice the range of $g(X)$ is finite, and so $g(X)$ is a discrete random variable; in particular the pmf of $g(X)$ is

$$p(x) = \begin{cases} 1/4 & \text{if } x = -1 \\ 3/4 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

If you feel that the example above is cheating a little bit since the function $g$ is not continuous, consider replacing the $g$ in the example by a constant function, say $g(x) = 1$. This $g$ is certainly continuous, but again produces a discrete random variable, although a very boring one.

Although composing a continuous random variable with a continuous function does not necessarily give us a new *continuous* random variable, there are a few cases where this is guaranteed.

**Theorem 10.1.**
*If $X : \Omega \to \mathbb{R}$ is a continuous random variable and $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing continuous function, then $g(X)$ is a continuous random variable.*

*Proof.*
Recall that a random variable is continuous if its cdf is continuous. Let $F$ be the cdf of the original random variable $X$, and let $G$ be the cdf of the new random variable $g(X)$. We want to show that

$G(x) = \Pr(g(X) \leq x)$ is a continuous function. Notice that since $g$ is strictly increasing, it is invertible. Thus $g(X) \leq x$ if and only if $X \leq g^{-1}(x)$. That is, $G(x) = F(g^{-1}(x))$. Now we use a slightly non-obvious fact: if $g$ is a strictly increasing continuous function, then its inverse is also continuous. This means $G$ is the composition of two continuous function, and so is continuous. □

If the function $g$ in Theorem 10.1 above is also differentiable, then we can in fact compute the pdf of $g(X)$ in terms of the pdf of $X$.

**Theorem 10.2.**
*If $X : \Omega \to \mathbb{R}$ is a continuous random variable with pdf $f(x)$ and if $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing differentiable function with $g'(x) > 0$ for all $x$, then $g(X)$ has pdf*

$$h(x) = \begin{cases} \frac{f(g^{-1}(x))}{g'(g^{-1}(x))} & \text{if } x \text{ is in the range of } g \\ 0 & \text{otherwise} \end{cases}$$

*Proof.*
Recall the pdf of a continuous random variable is the derivative of its cdf. From Theorem 10.1 we know the cdf of $g(X)$ is $G(x) = F(g^{-1}(x))$. Now we simply differentiate this to obtain

$$\begin{aligned} G'(x) &= F'(g^{-1}(x)) \cdot \frac{d}{dx} g^{-1}(x) \\ &= F'(g^{-1}(x)) \cdot \frac{1}{g'(g^{-1}(x))} \\ &= \frac{f(g^{-1}(x))}{g'(g^{-1}(x))}. \end{aligned}$$

Where above we used the the chain rule to differentiate $g^{-1}(x)$ as

follows:

$$\begin{aligned} & g(g^{-1}(x)) = x \\ \implies & \frac{d}{dx} g(g^{-1}(x)) = \frac{d}{dx} x \\ \implies & g'(g^{-1}(x)) \cdot \frac{d}{dx} g^{-1}(x) = 1 \\ \implies & \frac{d}{dx} g^{-1}(x) = \frac{1}{g'(g^{-1}(x))}. \end{aligned}$$

□

**Remark.**
Notice that while $g'(x) > 0$ for all $x$ certainly implies $g$ is strictly increasing, $g$ *can* be differentiable and strictly increasing without having $g'(x) > 0$ everywhere. For example, $g(x) = x^3$ is strictly increasing, though $g'(0) = 0$.

We need this stronger condition that $g'(x) > 0$ because we divide by $g'(g^{-1}(x))$ in the formula for the pdf of $g \circ X$ above.

**Example 10.9.**
Suppose $X$ is a continuous random variable with pdf

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Now suppose we compose $X$ with the function

$$g(x) = \begin{cases} 3x + 1 & \text{if } x < 0 \\ (x+1)^3 & \text{if } x \geq 0 \end{cases}$$

Notice the derivative of this function is

$$g'(x) = \begin{cases} 3 & \text{if } x < 0 \\ 3(x+1)^2 & \text{if } x \geq 0. \end{cases}$$

This derivative is obviously positive everywhere, so we can apply our theorem above to compute the pdf of $g \circ X$.

We will also need the inverse of $g$, but this is easy to compute. Notice that our function $g(x)$ transitions between two rules when $x = 0$, and $g(0) = 1$. This means the inverse will transition between two rules at $x = 1$.

$$g^{-1}(x) = \begin{cases} \frac{x-1}{3} & \text{if } x < 1 \\ \sqrt[3]{x} - 1 & \text{if } x \geq 1. \end{cases}$$

By Theorem 10.2, we can compute the pdf of $g \circ X$ as

$$\frac{f(g^{-1}(x))}{g'(g^{-1}(x))}.$$

Keeping in mind $f(x)$ is non-zero only when $0 \leq x \leq 2$, we see the function above is non-zero only when $0 \leq g^{-1}(x) \leq 2$.

Notice that $g^{-1}(x) < 0$ if $x < 1$, and $g^{-1}(x) > 2$ if $x > 27$. On the interval $1 \leq x \leq 27$ we have

$$\frac{f(g^{-1}(x))}{g'(g^{-1}(x))} = \frac{f(\sqrt[3]{x} - 1)}{g'(\sqrt[3]{x} - 1)} = \frac{3(\sqrt[3]{x} - 1)^2}{24\sqrt[3]{x^2}}$$

This means the pdf of $g \circ X$ is

$$h(x) = \begin{cases} \frac{3(\sqrt[3]{x}-1)^2}{24\sqrt[3]{x^2}} & \text{if } 1 \leq x \leq 27 \\ 0 & \text{otherwise} \end{cases}$$

Of course, it's not a big jump to modify the discussions above from increasing functions to decreasing functions, although there is one slightly subtle point. If $g(x)$ is increasing, then $g(X) \leq x$ if and only if $X \leq g^{-1}(x)$. If $g(x)$ is decreasing, however, then $g(X) \leq x$ if and only if $X \geq g^{-1}(x)$. This means if $G(x)$ is the cdf of $g \circ X$, then

$$G(x) = \Pr(g(X) \leq x) = \Pr(X \geq g^{-1}(x)) = 1 - \Pr(X < g^{-1}(x)) = 1 - F(g^{-1}(x)).$$

Differentiating both sides of $G(x) = 1 - F(g^{-1}(x))$ tells us the pdf is

$$G'(x) = -F'(g^{-1}(x)) \cdot \frac{d}{dx} g^{-1}(x) = \frac{-f(g^{-1}(x))}{g'(g^{-1}(x))}.$$

Notice that since $g$ is decreasing, $g'(x) < 0$ so the negative that appears actually insures our pdf is positive. These observations prove the following corollary.

**Corollary 10.3.**
*If $X : \Omega \to \mathbb{R}$ is a continuous random variable with pdf $f(x)$ and if $g : \mathbb{R} \to \mathbb{R}$ is a strictly decreasing differentiable function with $g'(x) < 0$ for all $x$, then $g(X)$ has pdf*

$$h(x) = \begin{cases} \left| \frac{f(g^{-1}(x))}{g'(g^{-1}(x))} \right| & \text{if } x \text{ is in the range of } g \\ 0 & \text{otherwise} \end{cases}$$

More generally, if we can break the real line up into a collection of intervals where $g(x)$ is strictly increasing or strictly decreasing with $g'(x) > 0$ or $g'(x) < 0$, then on each of those segments we can compute the pdf of $g \circ X$ as on each interval with the formulas from Theorem 10.2 and Corollary 10.3. Although now we have to worry about the fact that our function $g$ is not bijective, and so we actually need to sum up the values computed on increasing intervals with Theorem 10.2 and on decreasing intervals with Corollary 10.3. That is, if $g$ is a differentiable function whose derivative has finitely-many roots, then the pdf of $g(X)$ is

$$h(x) = \sum_{y \in g^{-1}(\{x\})} \left| \frac{f(y)}{g'(y)} \right|.$$

where we adopt the convention the sum is zero if $f^{-1}(\{x\}) = \emptyset$, and we only sum at points where $g'(y) \neq 0$.

(Notice if $f^{-1}(\{x\})$ contains only one point, then we get back the equations from Theorem 10.2 and Corollary 10.3.)

You may worry that we can't define the pdf at a few particular points (those points where $g'(x) = 0$), but this isn't really a big deal. The only thing we ever do with pdf's is integrate them, so if there's a few places the pdf isn't define it doesn't really matter because those points won't affect the

integral. More precisely, if we know that $g$ is differentiable and for every $x_0$ with $g'(x_0) = 0$ there exists some interval $(x_0 - \delta, x_0 + \delta)$ where $g'(x) \neq 0$ for all $x$ in this interval, then our pdf $h(x)$ will be defined at "enough" points for the integral of $h(x)$ to be well-defined.

**Example 10.10.**
Suppose $X$ is a continuous random variable with pdf

$$f(x) = \begin{cases} \frac{3x^2}{2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Now suppose we compose $X$ with the function $g(x) = x^2$. Notice that $g$ is strictly decreasing on $(-\infty, 0)$ and strictly increasing on $(0, \infty)$.

Notice that for each $x \in (-\infty, 0)$, $g^{-1}(x)$ is not defined since squaring a negative number produces a positive number. This means the pdf $h(x) = 0$ on $(-\infty, 0)$.

It's also clear that $h(x) = 0$ on $(1, \infty)$ since the outputs of the original random variable $X$ are between $-1$ and 1, and $g$ squares those values: $g \circ X$ can only take on values in $(0, 1)$.

If $x \in (0, 1)$, then $x$ has two preimages under $g$: $\pm\sqrt{x}$. At such point a point $h(x)$ is equal to

$$\begin{aligned} &\left|\frac{f(-\sqrt{x})}{g'(-\sqrt{x})}\right| + \left|\frac{f(\sqrt{x})}{g'(\sqrt{x})}\right| \\ =&\left|\frac{3x}{4(-\sqrt{x})}\right| + \left|\frac{3x}{4\sqrt{x}}\right| \\ =&\frac{3x}{2\sqrt{x}} \end{aligned}$$

The pdf $h(x)$ of $g \circ X$ is thus

$$h(x) = \begin{cases} \frac{3x}{2\sqrt{x}} & \text{if } 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

It's nice to know there are formulas we can use the compute the pdf of a composition of a continuous random variable with a (nice enough) continuous function, but these computations can pretty tedious. Luckily, if

all we're concerned with is the expected value of the composition, we can forego computing the pdf.

**Theorem 10.4.**
*If $X$ is a continuous random variable with pdf $f$, and if $g : \mathbb{R} \to \mathbb{R}$ is a function such that $g \circ X$ is a continuous random variable, then*

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

*Proof.*
We'll prove this only in the case where $g$ is differentiable and $g'(x) > 0$ everywhere. The proof when $g'(x) < 0$ is basically identical, and the proof when $g'(x) = 0$ only at isolated points follows from breaking up into intervals where the function is strictly increasing or strictly decreasing, it's just a little tedious to write down precisely.

For any interval $(-b, b)$, performing the substitution $u = g(x)$, $du = g'(x)dx$ we may write $x = g^{-1}(u)$ and $dx = \frac{du}{g'(x)} = \frac{du}{g'(g^{-1}(u))}$ and thus

$$\int_{-b}^{b} g(x)f(x)\,dx = \int_{g(-b)}^{g(b)} u\frac{f(g^{-1}(u))}{g'(g^{-1}(u))}\,du.$$

Of course, the $u$ on the right-hand side above is a "dummy variable," and we can rewrite it as $x$. Before doing that, though, notice the pdf of $g \circ X$ has appeared in our integral (computing the pdf using Theorem 10.2). That is,

$$\int_{-b}^{b} g(x)f(x)\,dx = \int_{g(-b)}^{g(b)} x\frac{f(g^{-1}(x))}{g'(g^{-1}(x))}\,dx = \int_{g(-b)}^{g(b)} xh(x)\,dx.$$

Taking the limit as $b$ goes to infinity we have

$$\int_{-\infty}^{\infty} g(x)f(x)\,dx = \lim_{b\to\infty}\int_{g(-b)}^{g(b)} xh(x)\,dx.$$

Notice the limits $\lim_{b\to\infty} g(b)$ and $\lim_{b\to\infty} g(-b)$ exist (or are $\infty$ and $-\infty$, respectively) since $g$ is increasing. In the event these limits are

not $\pm\infty$, we define $h(x) = 0$ for all $x < \lim_{b\to\infty} g(-b)$ and all $x > \lim_{b\to\infty} g(b)$ and we have

$$\int_{-\infty}^{\infty} g(x)f(x)\,dx = \int_{-\infty}^{\infty} xh(x)\,dx = \mathbb{E}[X].$$

□

Let's now compute the expected values of the random variable from Example 10.10 both using the pdf and the formula from Theorem 10.4 and see they are the same values.

**Example 10.11.**
Using the pdf

$$h(x) = \begin{cases} \frac{3x}{2\sqrt{x}} & \text{if } 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

from Example 10.10, we calculate the expected value

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} xh(x)\,dx \\ &= \int_0^1 \frac{3x^2}{2\sqrt{x}}\,dx \\ &= \frac{3}{2}\int_0^1 x^{3/2}\,dx \\ &= \frac{3}{2}\cdot\frac{2}{5}x^{5/2}\Big|_0^1 \\ &= \frac{3}{5}. \end{aligned}$$

Using Theorem 10.4 to compute the expected value with the original pdf

$$f(x) = \begin{cases} \frac{3x^2}{2} & \text{if } -1 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

gives

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\
&= \int_{-1}^{1} x^2 \cdot \frac{3x^2}{2}\,dx \\
&= \frac{3}{2}\int_{-1}^{1} x^4\,dx \\
&= \frac{3}{2}\cdot\frac{1}{5}x^5\Big|_{-1}^{1} \\
&= \frac{3}{10}\left(1^5 - (-1)^5\right) \\
&= \frac{3}{10}\cdot 2 \\
&= \frac{3}{5}.
\end{aligned}$$

## 10.6 Variance and standard deviation

Variance and standard deviation are defined for continuous random variables in exactly the same way they are defined for discrete random variables: if $X$ is a continuous random variable, then the ***variance*** of $X$ is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2],$$

and the standard deviation is

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

Though the definitions are the same, the actual computation is of course different since we must integrate to find the expected value above. Aside from the fact the computations are done differently, everything we know about variance and standard deviation for discrete random variables carries over to continuous random variables.

**Exercise 10.1.**

Show that if $X$ is a continuous random variable, then

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

**Example 10.12.**
Recall our running example of the continuous random variable $X$ with pdf

$$f(x) = \begin{cases} \frac{3x^2}{8} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We computed earlier that the expected value of this random variable was $3/2$. We can thus compute the variance as

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,dx - \left(\frac{3}{2}\right)^2 \\ &= \int_0^2 \frac{3x^4}{8}\,dx - \frac{9}{4} \\ &= \frac{3x^5}{40}\bigg|_0^2 - \frac{9}{4} \\ &= \frac{96}{40} - \frac{9}{4} \\ &= \frac{96-90}{40} \\ &= \frac{6}{40} \\ &= \frac{3}{20}. \end{aligned}$$

The standard deviation is $\sqrt{3/20}$.

## 10.7 Practice problems

**Problem 10.1.**
Suppose the cumulative distribution function of a continuous random variable $X$ is the following:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{x^2}{16} & \text{if } 0 \leq x < 2 \\ \frac{1}{4} & \text{if } 2 \leq x < 4 \\ \frac{x-3}{4} & \text{if } 4 \leq x < 7 \\ 1 & \text{if } x \geq 7 \end{cases}$$

What is the probability density function, $f(x)$, of this random variable?

**Problem 10.2.**
Let $X$ be a continuous random variable with the following probability density function,

$$f(x) = \begin{cases} \frac{10}{x^2} & \text{if } x > 10 \\ 0 & \text{otherwise} \end{cases}$$

(a) Verify that $f(x)$ is a probability density function.

(b) Compute the cumulative distribution function, $F(x)$, of $X$.

**Problem 10.3.**
Suppose $X$ is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} k(4x - 2x^2) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

for some constant $k$.

(a) What value of $k$ makes $f(x)$ a probability density function?

(b) What is $P(1/2 < X < 3/2)$?

(c) What is $\mathbb{E}[X]$?

**Problem 10.4.**
Suppose the cumulative distribution function of a continuous random variable $X$ is the following:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{x^2}{16} & \text{if } 0 \leq x < 2 \\ \frac{1}{4} & \text{if } 2 \leq x < 4 \\ \frac{x-3}{4} & \text{if } 4 \leq x < 7 \\ 1 & \text{if } x \geq 7 \end{cases}$$

What is the probability density function, $f(x)$, of this random variable?

**Problem 10.5.**
Suppose $X$ is a continuous random variable with the following pdf:

$$f(x) = \begin{cases} 4x^3 & \text{if } 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Compute $P(X \leq 2/3 \,\big|\, X \geq 1/2)$.

**Problem 10.6.**
Suppose $X$ is a continuous random variable with the following probability density function:

$$f(x) = \begin{cases} 2(1-x) & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Compute the median of $X$.

(b) Compute the variance of $X$.

# 11 Families of Continuous Random Variables

*Wahrlich es ist nicht das Wissen, sondern das Lernen, nicht das Besitzen sondern das Erwerben, nicht das Da-Seyn, sondern das Hinkommen, was den grössten Genuss gewährt.*
It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment.

Carl Friedrich Gauss

Just as we have families of discrete random variables which come up over and over again, like the Bernoulli, binomial, and Poisson, we also have families of continuous random variables. In this chapter we discuss three particular families. The uniform random variable is perhaps the simplest possible continuous random variable. The exponential random variable corresponds to times between independent random occurences, and is closely related to the Poisson random variables we discuss earlier. The normal random variable is in some sense the grand daddy of all continuous random variables. Its definition is considerably more complicated than the previous random variables we've discussed, but as we'll see later the normal random variables model many natural real-world phenomena.

## 11.1 Uniform

Intuitively, a uniform random variable represents a choice of a random number in some interval $[A, B]$ where all numbers are equally likely to be selected. This is only an intuitive description, however, since for any continuous random variable the probability of selecting any one particular value is always zero. To be more precise, a uniform random variable means the probability of selecting a value inside a subinterval $[C, D]$ of $[A, B]$ depends only on the size of $[C, D]$: all intervals of the same size are equally likely to contain the randomly selected point.

From this description we can determine the pdf of a uniform random variable. For the moment let's restrict ourselves to the interval $[0, 1]$. Our

goal is to find the pdf, or equivalently cdf, of a random variable with the following properties:

- The pdf is non-zero only on $[0,1]$. Equivalently, for the cdf we require $F(0) = 0$ and $F(1) = 1$.
- For any two intervals $[a,b]$ and $[c,d]$ contained in $[0,1]$ such that $d-c = b-a$, we require

$$F(b) - F(a) = \int_a^b f(x)\,dx = \int_c^d f(x)\,dx = F(d) - F(c).$$

As $[0,1] = [0,1/2] \cup [1/2,1]$, if we let $x = F(1/2) - F(0) = F(1) - F(1/2)$, then $2x = 1$. Note, though, $x = F(1/2)$ by the first equation. Similarly, writing $[0,1] = [0,1/3] \cup [1/3,2/3] \cup [2/3,1]$ shows that each interval has probability $1/3$ since the three intervals all have the same size and the sum of probabilities adds to 1. This means $F(1/3) = 1/3$ as $\Pr(X \in [0,1/3]) = F(1/3) - F(0)$. Notice this also means $F(2/3) = 2/3$ since

$$\begin{aligned} F(2/3) =& \Pr(X \in [0,2/3]) \\ =& \Pr(X \in [0,1/3] \cup [1/3,2/3]) \\ =& 2F(1/3). \end{aligned}$$

More generally, for any rational number $p/q$ in $[0,1]$ we see $F(p/q) = p/q$, thus $F(x) = x$ for all rational numbers $x$ in $[0,1]$, and by continuity $F(x) = x$ for all real numbers in $x$.

That is, for the random variable taking on values in $[0,1]$ where all subintervals of the same size are equally likely, the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \le x \le 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Differentiating, we see the pdf is

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1. \end{cases}$$

We can of course define the function to be zero at $x = 0$ and $x = 1$ to make sure the function is defined everywhere and this does not affect any integrals we would calculate.

The continuous random variable $X$ defined with the cdf and pdf above is called the ***uniform random variable supported on*** $[0,1]$ and we write $X \sim \text{Uni}([0,1])$ to indicate this.

We can easily define this random variable on any arbitrary interval $[A,B]$ without repeating the entire discussion above by doing a change of coordinates. That is, we can calculate probabilities by integrating the function $f(x)$ above over subintervals of $[0,1]$. If we would like to do the calculation over an interval $[A,B]$, we will just perform the $u$-substitution which transforms $[0,1]$ to $[A,B]$. This means we a linear function which sends 0 to $A$ and sends 1 to $B$, which is given by $u = (B-A)x + A$, so $du = (B-A)dx$ or $dx = \frac{1}{B-A}du$. Since $f(x)$ is 1 on $[0,1]$, the transformed pdf will be $\frac{1}{B-A}$ on $[A,B]$.

That is, a random variable $X$ is called the ***uniform random variable supported on*** $[A,B]$, denoted $X \sim \text{Uni}([A,B])$, if the pdf of $X$ is

$$f(x) = \begin{cases} 0 & \text{if } x < A \\ \frac{1}{B-A} & \text{if } A \leq x \leq B \\ 0 & \text{if } x > B \end{cases}$$

Once the pdf is known, it is of course a simple calculation to compute the cdf. If $X \sim \text{Uni}([A,B])$, then the cdf of $X$ is

$$F(x) = \begin{cases} 0 & \text{if } x < A \\ \frac{x-A}{B-A} & \text{if } A \leq x \leq B \\ 1 & \text{if } x > B \end{cases}$$

The uniform random variable is particularly simple, so we leave the verification of basic properties as exercises.

**Exercise 11.1.**
Let $X \sim \text{Uni}([A,B])$ and compute $\mathbb{E}[X]$.

**Exercise 11.2.**
Let $X \sim \text{Uni}([A,B])$ and compute $\text{Var}(X)$.

**Exercise 11.3.**
For each $p \in (0,1)$, compute the $(100 \cdot p)$-th percentile of $X \sim \text{Uni}([A,B])$.

## 11.2 Exponential

The second family of continuous random variables we will discuss are the exponential random variables, which are closely related to the discrete Poisson random variables we discussed earlier. We will first just give the definition and mention some basic properties of exponential random variables, and then describe the relationship to the Poisson.

We say a continuous random variable $X$ is an ***exponential random variable with parameter*** $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if the pdf of $X$ is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

From this we can easily compute the cdf of $X$. Notice first that since $f(x) = 0$ for $x < 0$, the cdf $F$ satisfies $F(x) = 0$ for $x < 0$ as well. For $x \geq 0$ we compute

$$\begin{aligned} F(x) &= \Pr(X \leq x) \\ &= \Pr(0 \leq X \leq x) \\ &= \int_0^x \lambda e^{-\lambda t}\, dt \end{aligned}$$

Performing the substitution $u = -\lambda t$, $du = -\lambda dt$ we have

$$\begin{aligned} F(x) &= \int_0^x \lambda e^{-\lambda t}\, dt \\ &= \int_0^{-\lambda x} -e^u\, du \\ &= \int_{-\lambda x}^0 e^u du \\ &= e^u \Big|_{-\lambda x}^0 \\ &= e^0 - e^{-\lambda x} \\ &= 1 - e^{-\lambda x} \end{aligned}$$

The cdf is thus

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0. \end{cases}$$

We can now prove one of the most important properties of exponential random variables, which is that they are *memoryless*. What this means is that if we already know $X > s$, then the probability $X > t$ (for some $t > s$) is equal to the probability $X > t - s$. That is, the random variable is always, continually "resetting" itself. For example, memorylessness means $\Pr(X > 5|X > 2) = \Pr(X > 3)$. For example, perhaps you measure the amount of time between random occurences of some event – say the time between consecutive cars taking an exit on the highway. If you know that two minutes have already passed since the last car took the exit, the probability it will be more than five minutes between cars is the same as the probability of waiting another three minutes. Because the cars are assumed to be independent of one another, knowledge of when the last car took an exit tells you nothing about when the next car will take the exit, and so you can imagine that you are continually restarting the experiment at every instance of time.

**Theorem 11.1.**
*Exponential random variables are memoryless. That is, if* $X \sim \text{Exp}(\lambda)$, *then for any* $t > s > 0$,

$$\Pr(X > t|X > s) = \Pr(X > t - s).$$

*Proof.*
First notice that

$$\Pr(X > x) = 1 - \Pr(X \leq x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}.$$

Now we simply compute $\Pr(X > t|X > s)$ using the definition of

conditional probability:

$$\begin{aligned}\Pr(X > t | X > s) &= \frac{\Pr([X > t] \cap [X > s])}{\Pr(X > s)} \\ &= \frac{\Pr(X > t)}{\Pr(X > s)} \\ &= \frac{e^{-\lambda t}}{e^{-\lambda s}} \\ &= e^{-\lambda t - (-\lambda s)} \\ &= e^{-\lambda(t-s)} \\ &= \Pr(X > t - s).\end{aligned}$$

□

As always, any time we have a random variable we might want to know its expected value and variance.

**Theorem 11.2.**
*If* $X \sim \text{Exp}(\lambda)$*, then* $\mathbb{E}[X] = 1/\lambda$ *and* $\text{Var}(X) = 1/\lambda^2$.

**Exercise 11.4.**
Prove Theorem 11.2.

## Relation to Poisson random variables

Recall that a Poisson random variable with parameter $\lambda$ counts the number of occurences of a random phenomenon in a given time interval, where the average number of occurences is known to be $\lambda$. For example, counting the number of fish that under a bridge over the course of an hour. The time between two consecutive such occurences (e.g., the time we have to wait after one fish swims by until the next fish swims by) is itself a random variable,

and in fact is precisely an exponential random variable with parameter the same $\lambda$ as in the Poisson random variable.

To prove that the ***interarrival times*** between random occurences counted by the Poisson are exponential random variables, we first need to recall that each Poisson random variable gives rise to a Poisson process. That is, if the number of occurences in a unit time interval is $\lambda$ on average, the number of occurences in a time interval of length $t$ is $\lambda t$ on average. For example, if on average the number of fish that swim by a bridge over an hour is 42, then the number of fish swimming by per three hours is on average $42 \cdot 3 = 126$, the number of fish swimming by every fifteen minutes (quarter of an hour) is on average $42 \cdot 1/4 = 10.5$, and so on.

Now let $N(t)$ be our Poisson process counting the number of occurences up to time $t$, so $N(t) \sim \text{Poisson}(\lambda t)$. For each $t$, let $T(t)$ denote the remaining time until the next random occurrence. To make this more explicit, let's consider an explicit example.

Suppose as above we're counting the number of fish swimming by a bridge, and let $t$ be measured in hours. Then $N(1)$ is the number of fish we've counted over the course of one hour, $N(2)$, is the number of fish we've counted over two hours, $N(2.75)$ is the number of fish over two hours and forty-five minutes, etc. For concreteness, suppose that a fish swims by at $t = 0.75$ (45 minutes after we start counting), and so far our count of fish is maybe 30: $N(0.75) = 30$. Now suppose no more fish swim by until time $t = 0.875$ (52.5 minutes after our count starts). That is, $N(t) = 30$ for $0.75 \leq t < 0.875$, but then $N(t) = 31$ at $t = 0.875$ and stays at 31 until another fish swims by, whenever that happens to be. The $T(t)$ measures the time between consecutive fish. For example, $T(0.75) = 0.125$ since at time $t = 0.75$ we don't see another fish until $t = 0.875$; similarly, $T(0.8) = 0.075$ since, again, we won't see another fish for 0.075 hours from the current time. Because the times that the fish swim by are random (we don't know when the next fish will swim by), $T(t)$ is a random variable, and our goal is to figure out what type of random variable it is. The claim we're trying to prove is that if we somehow know $N(t)$ is Poisson, then $T(t)$ must be exponential.

Suppose that we knew we had to wait some given amount of time, call it $\tau$, until the next random occurrence from time $t$. That is, suppose we knew $T(t) > \tau$. What does this tell us about our counting process, $N(t)$? Since we don't see another random occurence for another $\tau$ units of time, our count must remain the same. That is, if $T(t) > \tau$, then $N(t) = N(t + \tau)$. Conversely, if we somehow knew $N(t) = N(t + \tau)$, then the next random occurence doesn't occur before an additional $\tau$ units of time, and so $T(t) > \tau$. This is a convenient observation because it allows

us to translation statements about $T(t)$, which we don't yet know and are trying to understand, into statements about $N(t)$, which we are assuming is Poisson.

In terms of probability, the above observations justify the following string of equalities:

$$\Pr(T(t) \leq \tau) = 1 - \Pr(T(t) > \tau) = 1 - \Pr(N(t+\tau) = N(t))$$

So, determining the distribution of $T(t)$ is equivalent to determining the probability $N(t+\tau) = N(t)$, or equivalently the probability $N(t+\tau) - N(t) = 0$. Notice this means our Poisson process counted no additional random occurences over a time interval of length $\tau$. But the number of random occurences is itself a Poisson random variable with parameter $\lambda\tau$. That is, $N(t+\tau) - N(t) \sim \text{Poisson}(\lambda\tau)$, and we can easily calculate the probability this is zero:

$$\Pr(N(t+\tau) - N(t) = 0) = \frac{(\lambda\tau)^0}{0!}e^{-\lambda\tau}.$$

Plugging this into our string of equalities above we have

$$\Pr(T(t) \leq \tau) = 1 - e^{-\lambda\tau},$$

but this is exactly the cdf of an exponential random variable with parameter $\lambda$. That is, we have proven the following theorem.

**Theorem 11.3.**
*The interarrival times between random occurences counted by a Poisson random variable with parameter $\lambda$ is an exponential random variable with parameter $\lambda$.*

**Example 11.1.**
Suppose the number of potholes per mile of roadway in Indiana is modeled by a Poisson random variable with an average of three potholes per mile. What is the average distance between two consecutive potholes? What is the probability the distance between two consecutive potholes is at most half a mile? What is the probability the distance between two consecutive potholes is at least one mile?

In the notation above, $N(t)$ counts the number of potholes per

$t$ miles of roadway. We are told $N(1) \sim \text{Poisson}(3)$, so in general $N(t) \sim \text{Poisson}(3t)$. By Theorem 11.3, we know the distance between consecutive potholes is $D \sim \text{Exp}(3)$ (we'll call this $D$ instead of $T$ since here we are measuring distance instead of time). Hence the average distance between consecutive potholes is $\mathbb{E}[D] = 1/3$ by Theorem 11.2.

The probability the distance between consecutive potholes is at most half a mile is $\Pr(D \leq 1/2) = 1 - e^{-3 \cdot 1/2} \approx 0.7769$.

The probability the distance between consecutive potholes is at least one mile is $\Pr(D > 1/2) = e^{-3 \cdot 1} \approx 0.0498$.

## 11.3 Normal

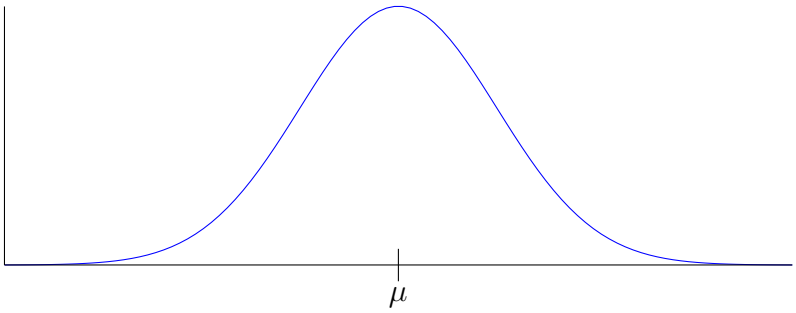
The ***normal random variables***, also sometimes called ***Gaussian random variables***, are some of the most important ones in all of probability and statistics, although the reasons for this will have to wait until we discuss jointly distributed random variables and the central limit theorem to see why this is the case, and now simply treat the normal random variables as another example of continuous random variables.

The formulas involved in defining the normal random variables below may look complicated, but they are important for many reasons. Many natural phenomena can be accurately modelled by normal random variables and there is some serious theory behind why this is the case. We're not quite ready for the general theorey just yet, so we'll take it on faith for the moment that normal random variables model real-world phenomena of interest.

Normal random variables depend on two parameters, $\mu$ and $\sigma$, which will turn out to be the expected value and standard deviation. We say that a continuous random variable $X$ is ***normally distributed*** with parameters $\mu$ and $\sigma > 0$, denoted $X \sim N(\mu, \sigma)$ if the pdf of $X$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

The graph of such a function this function has the form




(This function never actually touches the $x$-axis, but the $y$-values get so small that it appears to in the picture above.)

Note first that this is defined for all $x$. It's clear from the definition that $f(x) > 0$ for all $x$ since $\sigma$ is positive and the exponential function $e^x$ is always positive. It is not at all obvious that the $f(x)$ above will integrate to 1, however, nor is it even immediately obvious how to go about integrating the function above. It turns out we can actually verify the function above integrates to 1, though it requires some non-obvious trickery.

**Proposition 11.4.**
*For any real number $\mu$ and any $\sigma > 0$,*

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\, dx = 1.$$

*Proof.*
We'll first prove this in the special case when $\mu = 0$ and $\sigma = 1$, and then perform a $u$-substitution to transform any other choice of $\mu$ and $\sigma$ into the $\mu = 0$ and $\sigma = 1$ case.

When $\mu = 0$ and $\sigma = 1$, we are trying to evaluate the integral

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

Let's first notice that the integrand is an even function: the integrand evaluates to the same value for $x$ and $-x$ because of the squaring

involved. Thus we can rewrite the integral as

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx = 2\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx.$$

Now we write the integral as the square root of its square,

$$\begin{aligned} \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx &= \frac{2}{\sqrt{2\pi}} \sqrt{\left( \int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx \right)^2} \\ &= \frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx \int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx}. \end{aligned}$$

We now rewrite this as an iterated integral,

$$\frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx \int_{0}^{\infty} e^{\frac{-x^2}{2}}\, dx} = \frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} \int_{0}^{\infty} e^{\frac{-(x^2+y^2)}{2}}\, dy\, dx}.$$

Performing the substitution $u = \frac{y}{x}$ (so, $y = ux$ and $dy = u\, dx$) this becomes

$$\frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} \int_{0}^{\infty} e^{\frac{-(x^2+u^2x^2)}{2}} x\, du\, dx} = \frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} \int_{0}^{\infty} e^{\frac{-x^2(1+u^2)}{2}} x\, dx\, du}$$

We now integrate with respect to $x$ to obtain

$$\frac{2}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} \left( \frac{-1}{1+u^2} e^{\frac{-x^2(1+u^2)}{2}} \Big|_0^{\infty} \right) du}$$

Notice that as $x$ goes to infinity, the factor $e^{\frac{-x^2(1+u^2)}{2}}$ goes to zero and

so the integral becomes

$$\begin{aligned}\frac{2}{\sqrt{2\pi}}\sqrt{\int_0^\infty \frac{1}{1+u^2}\,du} &= \frac{2}{\sqrt{2\pi}}\sqrt{\arctan(u)\Big|_0^\infty} \\ &= \frac{2}{\sqrt{2\pi}}\sqrt{\pi/2} \\ &= \frac{\sqrt{2}}{\sqrt{\pi}}\cdot\frac{\sqrt{\pi}}{\sqrt{2}} \\ &= 1.\end{aligned}$$

In the computation above we supposed $\mu = 0$ and $\sigma = 1$. For any other choice of $\mu$ and $\sigma > 0$, we simply perform the $u$-substitution

$$u = \frac{x-\mu}{\sigma} \qquad du = \frac{1}{\sigma}dx$$

to obtain

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}\,du.$$

The latter integral is simply the integral we computed above, however, and so this is just 1. □

In fact, the $f(x)$ above has the odd property that its integral can not be written in any simpler form; such functions are sometimes called ***non-elementary***. That is, the fundamental theorem of calculus promises us that the $f(x)$ above has an antiderivative, but it's impossible to write the antiderivative of this function as anything simpler than

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)^2}{2\sigma^2}}\,dt.$$

As a consequence, we can't really directly evaluate this integral – i.e., we can't compute the probabilities given to us by normal random variables!

We can, however, numerically approximate these values. (Think of doing something like a Riemann sum with a very large number of very skinny rectangles.) This is extremely tedious to do by hand, but luckily there's a little trick we can use to make these computations a little more tractable. To explain the trick we need to discuss one special choice of $\mu$ and $\sigma$.

The ***standard normal random variable***, often denoted $Z$, is the normal random variable with parameters $\mu = 0$ and $\sigma = 1$, $Z \sim N(0,1)$,

and so has pdf

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

The cdf of the standard normal is often denoted $\Phi$:

$$\Phi(z) = \Pr(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx.$$

Though we can't directly compute $\Phi(z)$, several values of $\Phi(z)$ have been approximated numerically by other people nad these values can be looked up in a book or on a computer, and we can thus use those approximations of $\Phi(z)$ to estimate probabilities $\Pr(a \leq Z \leq b)$.

For instance, it's known that $\Phi(1.3) = \Pr(Z \leq 1.3) \approx 0.9032$ and $\Phi(-0.5) = \Pr(Z \leq -0.5) \approx 0.3085$. From this we can estimate $\Pr(-0.5 \leq Z \leq 1.3)$ as

$$\begin{aligned} \Pr(-0.5 \leq Z \leq 1.3) &= \Phi(1.3) - \Phi(-0.5) \\ &\approx 0.9032 - 0.3085 \\ &= 0.5947. \end{aligned}$$

Since we can only calculate integrals of the pdf of $Z$ approximately, you may think we can only approximate the expected value and variance of $Z$, but the presence of an extra $x$ in $\mathbb{E}[X]$ and an $x^2$ in $\mathrm{Var}(X)$ actually make the integrals easier because we can then use $u$-substitution and integration by parts.

**Theorem 11.5.**
*If $Z \sim N(0,1)$ is the standard normal, then $\mathbb{E}[Z] = 0$ and $\mathrm{Var}(Z) = 1$.*

*Proof.*
To compute the expected value we need to calculate

$$\mathbb{E}[Z] = \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \, dx.$$

Now let's notice that we can rewrite this integral as

$$\int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x e^{\frac{-x^2}{2}}\, dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} x e^{\frac{-x^2}{2}}\, dx +$$

Now perform the substitution $u = \frac{-x^2}{2}$, $du = -x\,dx$ in each integral to obtain

$$-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^u\, du + -\frac{1}{\sqrt{2\pi}} \int_{0}^{-\infty} e^u\, du$$

We can now flip the change the order of integration in the second

$$\begin{aligned} &-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^u\, du + -\frac{1}{\sqrt{2\pi}} \int_{0}^{-\infty} e^u\, du \\ =&-\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^u\, du + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^u\, du \end{aligned}$$

and obviously this cancels out to give zero.

Since $\mathbb{E}[Z] = 0$, we know $\mathrm{Var}(Z) = \mathbb{E}[Z^2]$ which we compute as

$$\int_{-\infty}^{\infty} x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

We again break the integral up into two parts,

$$\mathbb{E}[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x^2 \cdot e^{\frac{-x^2}{2}}\, dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} x^2 \cdot e^{\frac{-x^2}{2}}\, dx. \qquad (11.1)$$

Now we perform integration by parts on each of these. For the first integral we take

$$\begin{aligned} u &= x & dv &= x e^{-\frac{x^2}{2}}\, dx \\ du &= dx & v &= e^{-\frac{x^2}{2}} \end{aligned}$$

the first integral above then becomes

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x^2 \cdot e^{\frac{-x^2}{2}}\, dx + = x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{0} - \int_{-\infty}^{0} e^{-\frac{x^2}{2}}\, dx$$

Writing the first term as a limit and using l'Hôpital's rule we have

$$\lim_{x\to-\infty} \frac{x}{e^{-\frac{x^2}{2}}} = \lim_{x\to-\infty} \frac{1}{xe^{-\frac{x^2}{2}}} = 0.$$

The second term of the first integral we'll leave alone for the moment. Performing the same integration by parts and l'Hôpital's calculation to the second integral in Equation 11.1 above will likewise show that the first part of the integral (after rewriting with integration by parts) is zero, while the second integral remains.

Altogether, we now have

$$\mathbb{E}[Z^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{x^2}{2}}\, dx + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\, dx$$

Notice this is simply the integral of the pdf for the standard normal, and so equals one. □

## Transforming normal random variables

In the above we placed treated the standard random normal $Z \sim N(0,1)$ as special: we gave it a special symbol, $Z$; we gave its cdf a special name, $\Phi$; and in showing the pdf of a normal random variable integrated to 1, we first proved this for $Z$. There's not really anything magical about $Z$ versus any other normal random variable, but it is often convenient to eliminate $\mu$ and $\sigma$ from our calculations by assuming they are 0 and 1, respectively. More importantly, we can transform *any* normal random variable $X \sim N(\mu, \sigma)$. The trick for doing this is hinted at in the proof of Proposition 11.4, but now we make it precise.

**Proposition 11.6.**
*If $X \sim N(\mu, \sigma)$, then $\frac{X-\mu}{\sigma} = Z$. That is, for any $a$ and $b$,*

$$\Pr(a \le X \le b) = \Pr\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right) - \Phi\left(\frac{b-\mu}{\sigma}\right).$$

We'll prove Proposition 11.6 in just a moment, but first let's think about

what the proposition tells us. This proposition says that if we have any normal, we can do a simple manipulation to get a standard normal. In particular, if we are able to compute (or look up) values of $\Phi$ in a book or on a computer, then we can use that information to calculate probabilities for any other normal random variable. So, in some way, the only normal random variable we really need to know how to work with is the standard normal, since we can transform any other normal into the standard normal.

*Proof of Proposition 11.6.*
Performing the $u$-substitution,

$$u = \frac{x-\mu}{\sigma} \qquad du = \frac{1}{\sigma}dx$$

the integral

$$\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}} dx$$

becomes

$$\int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$$

but this is precisely the integral of the pdf of the standard normal from $\frac{a-\mu}{\sigma}$ to $\frac{b-\mu}{\sigma}$. □

**Example 11.2.**
If $X \sim N(12, 3)$, what is $\Pr(8 \leq X \leq 13)$?

By Proposition 11.6, we have

$$\Pr(8 \leq X \leq 13) = \Pr\left(\frac{8-12}{3} \leq Z \leq \frac{13-12}{3}\right) = \Pr\left(\frac{-4}{3} \leq Z \leq \frac{1}{3}\right).$$

This probability is of course given by $\Phi\left(1/3\right) - \Phi\left(-4/3\right)$ which we can look up are is approximately $0.631 - 0.091 = 0.54$, and so $\Pr(8 \leq X \leq 13) \approx 0.54$.

We can also go backwards to convert $Z$ values into $X$ values, as shown

in the following example.

**Example 11.3.**
Given $\Pr(Z \leq 2) \approx 0.977$ and $X \sim N(5, 2)$, for what value of $x$ do we have $\Pr(X \leq x) \approx 0.977$?

As

$$\Pr(X \leq x) = \Pr\left(\frac{X-5}{2} \leq \frac{x-5}{2}\right) = \Pr\left(Z \leq \frac{x-5}{2}\right)$$

we want to find the value of $x$ such that $\frac{x-5}{2} = 2$, since we know $\Pr(Z \leq 2) \approx 0.977$. Of course this is a simple algebra problem, and solving for $x$ gives $x = 9$.

Thus if $X \sim N(5, 2)$, then $\Pr(X \leq 9) \approx 0.977$.

In general, if $\Pr(Z \leq z) = p$, then for the random variable $X \sim N(\mu, \sigma)$ we have $\Pr(X \leq x) = p$ when $x$ solves the equation $\frac{x-\mu}{\sigma} = z$; i.e., $x = z\sigma + \mu$. That is, $\Pr(Z \leq z) = \Pr(X \leq z\sigma + \mu)$.

**Example 11.4.**
Suppose scores on an IQ test are normally distributed with mean 100 and standard deviation 15. What is the 90-th percentile of these IQ scores?

Here, our random variable $X$ is an IQ score and we're told $X \sim N(100, 15)$. We want to find the value of $\eta$ such that $\Pr(X \leq \eta) = 0.9$. Notice that it suffices for us to find the 90-th percentile for the standard normal, since we can convert $X$ into $Z$ and vice versa. Looking up that $\Pr(Z \leq 1.282) = 0.9$, the above tells us $\Pr(X \leq 1.232 \cdot 15 + 100) = 0.9$ and so $\Pr(X \leq 119.23) = 0.9$. That is, the 90th percentile on these IQ tests is 119.23.

We can now easily extend Theorem 11.5 to compute the expected value and variance of any normal random variable by using our knowledge of the standard normal random variable and the transformations mentioned above.

**Theorem 11.7.**
*If* $X \sim N(\mu, \sigma)$, *then* $\mathbb{E}[X] = \mu$ *and* $\mathrm{Var}(X) = \sigma^2$.

*Proof.*
Recall that for any $m$ and $b$, $\mathbb{E}[mX + b] = m\mathbb{E}[X] + b$. Thus

$$\mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \mathbb{E}\left[\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right] = \frac{1}{\sigma}\mathbb{E}[X] - \frac{\mu}{\sigma} = \frac{\mathbb{E}[X]-\mu}{\sigma}.$$

Since $\frac{X-\mu}{\sigma} = Z$ and $\mathbb{E}[Z] = 0$, however, we have $\frac{\mathbb{E}[X]-\mu}{\sigma} = 0$ and solving for $\mathbb{E}[X]$ gives $\mu$.

For any $m$ and $b$, $\mathrm{Var}(mX + b) = m^2\mathrm{Var}(X)$. Now we simply note

$$1 = \mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X),$$

and solving for $\mathrm{Var}(X)$ gives $\sigma^2$. □

## 11.4 Practice problems

**Problem 11.1.**
Suppose that $X$ is a uniformly distributed continuous random variable on the interval $[-1, 1]$. Consider the random variable obtained by taking the absolute value, $|X|$.

(a) What is $P(|X| > 1/2)$?

(b) What is the cumulative distribution function of $|X|$?

**Problem 11.2.**
Suppose the number of automobile accidents in Monroe county each month is modeled by a normal random variable with a mean of 45 accidents per month and a standard deviation of 10. What is the probability there are more than sixty accidents in a given month?

**Problem 11.3.**
Suppose weights of newborn babies in the United States is normally distributed with an average of 8 pounds and standard deviation of 0.5 pounds. Estimate the 70-th percentile of these weights.

# 12 Jointly Distributed Random Variables

*Probability is too important to be left to the experts.*

RICHARD HAMMING

In this chapter we start working on making the transition from probability to statistics. We extend our theory of random variables developed thus far to work with several random variables at once. This is necessary for statistics because later we will view the data points obtained in a sample as individual random variables, and will want to work with all of the random variables simultaneously.

## 12.1 Joint discrete random variables

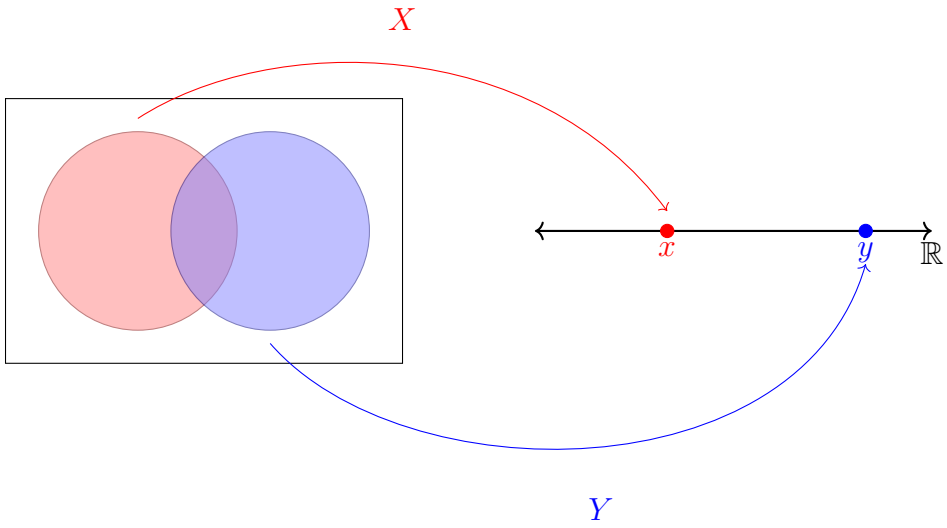
Many times we have several random variables that we are simultaneously interested in. For example, an insurance company may be interested in several different quantities associated to an automobile accident, such as the number of cars involved; the cost of damage to the cars; medical costs of passengers; the age of drives involved in the accident; and the number of occupants in each car. Each of these quantities is a different random variable associated to one accident, and the insurance company might be interested in each one, and also in any relationships between those pieces of data. For example, the age of the drive might be related to the number of occupants: young drivers might be more likely to ride around with several friends, while older drivers might primarily drive alone.

In general, if $X$ and $Y$ are two random variables defined on the same sample space $\Omega$, then we might be interested in the probability that $X$ takes on one value while simultaneously $Y$ takes on another value. If $X$ and $Y$ are both discrete random variables, then we define the ***joint probability mass function*** of $X$ and $Y$ as the following function of two variables,

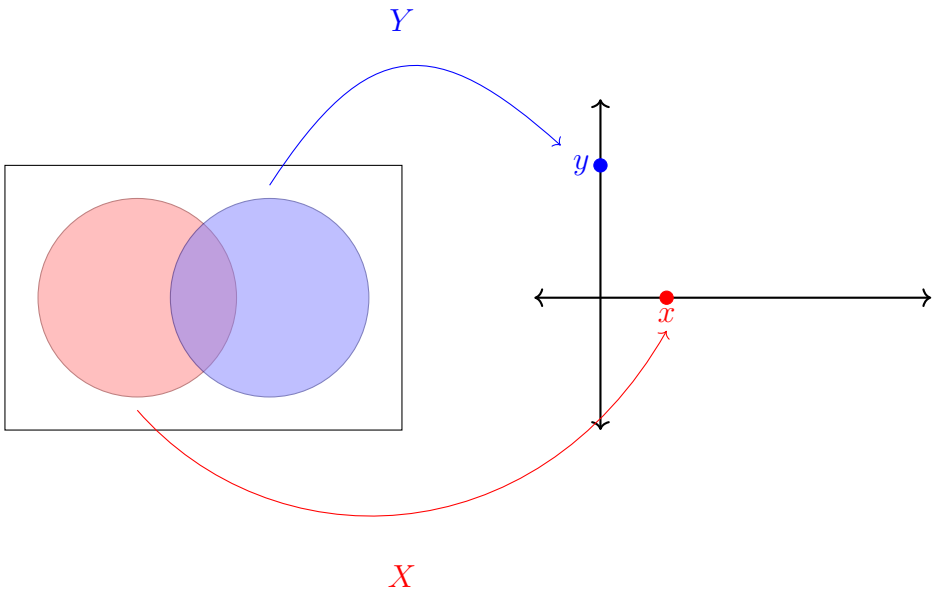
$$p(x, y) = \Pr(X = x \text{ and } Y = y) = \Pr(X^{-1}(\{x\}) \cap Y^{-1}(\{y\})).$$

In the schematic picture below, the red oval represents all of the points which the random variable $X$ maps to $x$, and the blue oval represents all

of the points which $Y$ maps to $y$. If we want both of these events to take place, then we are interested in the intersection of the two events.




Let's notice that we could think of $X$ and $Y$ as giving us the $(x, y)$-coordinates of a point in the plane.




More generally, we may want to know the probability $(X, Y)$ gives us a point inside some region $E$ of the plane, instead of a particular point. If the joint probability mass function is know, however, we can compute the probability $(X, Y) \in E$ by summing up the probabilities for all $(x, y)$ points

inside $E$:

$$\Pr((X,Y)\in E)=\sum_{(x,y)\in E} p(x,y).$$

For instance, suppose $p(x,y)$ is the joint pmf for two random variables given by the table below.

| $x$ \ $y$ | $1/4$ | $1/2$ | $3/4$ | $1$ |
|---|---|---|---|---|
| 1 | 0.02 | 0.13 | 0.07 | 0.03 |
| 2 | 0.1 | 0 | 0.02 | 0.005 |
| 3 | 0.2 | 0.15 | 0.01 | 0.015 |
| 4 | 0.05 | 0.05 | 0.025 | 0 |
| 5 | 0 | 0 | 0.125 | 0 |

That is, this table tells us

$$\Pr(X=2,Y=3/4)=p(2,3/4)=0.02.$$

Implicitly, for any $(x,y)$-point not in the table above, $p(x,y)=0$. For example, $p(1,7/8)=0$.

Given a region in the plane such as $E=[2,4]\times[1/2,1]$ (notice this means all of the $(x,y)$ points where $2\le x\le 4$ and $1/2\le y\le 1$), we compute the probability $(X,Y)\in E$ by summing up $p(x,y)$ for all $(x,y)$ values in our region. Of course, $p(x,y)$ will be zero for "most" of these points, so we only need to worry about summing over the points in the table above. In this particular case we have

$$\begin{aligned}&\Pr((X,Y)\in E)\\ =&p(2,1/2)+p(2,3/4)+p(2,1)+\\ &p(3,1/2)+p(3,3/4)+p(3,1)+\\ &p(4,1/2)+p(4,3/4)+p(4,1)\\ =&0.275\end{aligned}$$

Notice that we can recover the pmf's of the original random variable $X$ and $Y$ from the joint pmf. In this context, the pmf of a single random variable is called a ***marginal pmf***, and is denoted $p_X(x)$ or $p_Y(y)$ depending on whether we're computing the pmf of $X$ or of $Y$. We compute these marginal pmf's by summing over all choices of the other variable. That is,

$$p_X(x)=\sum_{y\in\mathbb{R}} p(x,y)$$
$$p_Y(y)=\sum_{x\in\mathbb{R}} p(x,y).$$

What's happening here is we're saying we want to find the probability $X = x$, regardless of what $Y$ is. Since the pmf tells us the probability that $X = x$ and $Y = y$, but we don't care about $Y$, we look at all possible values of $Y$ for our fixed value of $X$, and this tells us $p_X(x)$. For the joint pmf in the table above this gives us

$$p_X(x) = \begin{cases} 0.25 & \text{if } x = 1 \\ 0.125 & \text{if } x = 2 \\ 0.375 & \text{if } x = 3 \\ 0.125 & \text{if } x = 4 \\ 0.125 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 0.37 & \text{if } y = 1/4 \\ 0.33 & \text{if } y = 1/2 \\ 0.25 & \text{if } y = 3/4 \\ 0.05 & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

As $p(x, y)$ represents probabilities, there are a few obvious properties that must be satisfied:

1. $0 \leq p(x, y) \leq 1$ for all $(x, y)$, and
2. $\displaystyle\sum_{(x,y)\in\mathbb{R}^2} p(x, y) = 1.$

## 12.2 Joint continuous random variables

If $X$ and $Y$ are continuous random variables, their ***joint probability density function*** is the function of two variables $f : \mathbb{R}^2 \to \mathbb{R}$ that satisfies the following property: for any $E \subseteq \mathbb{R}$,

$$\Pr((X, Y) \in E) = \iint_E f(x, y)\, dA.$$

Geometrically, this means the probability above is given by the volume between the surface $z = f(x, y)$ and the region $E$ in the $xy$-plane.

**Remark.**
If you have not seen integration in several variables before, or if you need a refresher, the details are typed up in detail in Appendix A. The end result of this, however, is that if you can integrate in one variable, then you can integrate in two variables; just integrate one variable at a time.

In the simplest situations when the region $E$ in the plane is a rectangle such as $E = [a, b] \times [c, d]$, then the double integral $\iint_E f(x, y)\, dA$ is computed as either of the ***iterated integrals***:

$$\iint_E f(x, y)\, dA = \int_a^b \int_c^d f(x, y)\, dy\, dx = \int_c^d \int_a^b f(x, y)\, dx\, dy.$$

We compute these integrals from the inside out, treating one of the variables as a constant. For example, the integral

$$\int_0^1 \int_2^3 x^2 y\, dx\, dy$$

is computed by treating $y$ as a constant in the inner most integral to obtain

$$\int_0^1 \int_2^3 x^2 y\, dx\, dy = \int_0^1 \frac{x^3 y}{3}\bigg|_2^3 dy = \int_0^1 \frac{19y}{3}\, dy$$

Of course, at this point this is just a "normal" integral of one variable which we compute as

$$\int_0^1 \frac{19y}{3}\, dy = \frac{19y^2}{6}\bigg|_0^1 = \frac{19}{6}.$$

For more information, see the appendix at the end of these notes.

**Example 12.1.**

Suppose the joint pdf of two continuous random variables $X$ and $Y$ is

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Then the probability that $X$ is between 0 and $1/2$ while $Y$ is simultaneously between $1/4$ and $3/4$ is

$$\begin{aligned}
\Pr((X,Y) \in [0,1/2] \times [1/4,3/4]) &= \iint_{[0,1/2]\times[1/4,3/4]} \left(x + \frac{3}{2}y^2\right) dA \\
&= \int_0^{1/2} \int_{1/4}^{3/4} \left(x + \frac{3}{2}y^2\right) dy\, dx \\
&= \int_0^{1/2} \left(xy + \frac{y^3}{2}\right)\Bigg|_{1/4}^{3/4} dx \\
&= \int_0^{1/2} \left(\frac{3}{4}x + \frac{(3/4)^3}{2} - \left(\frac{1}{4}x + \frac{(1/4)^3}{2}\right)\right) dx \\
&= \int_0^{1/2} \left(\frac{1}{2}x + \frac{13}{192}\right) dx \\
&= \left(\frac{x^2}{4} - \frac{13x}{192}\right)\Bigg|_0^{1/2} \\
&= \frac{1}{16} - \frac{13}{384} \\
&= \frac{24-13}{384} \\
&= \frac{11}{384} \\
&\approx 0.0287
\end{aligned}$$

Just as the joint pmf of two discrete random variables satisfies some obvious properties, so does the joint pdf of two continuous random variables:

1. $0 \le f(x,y) \le 1$ for all $(x,y) \in \mathbb{R}^2$, and
2. $\iint_{\mathbb{R}^2} f(x,y)\, dA = 1$.

We can also recover the ***marginal pdf*** of each of the random variables

by simply integrating out the other random variable:

$$p_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$
$$p_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx.$$

**Example 12.2.**
If the joint pdf of two continuous random variables is

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

then the marginal pdf's are

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \begin{cases} \int_0^1 \left(x + \frac{3}{2}y^2\right)\,dy & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \left(xy + \frac{y^3}{6}\right)\Big|_0^1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} x + \frac{1}{6} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$\begin{aligned} p_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\ &= \begin{cases} \int_0^1 \left(x + \frac{3}{2}y^2\right)\,dx & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \left(\frac{x^2}{2} + \frac{3}{2}xy^2\right)\Big|_0^1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{1}{2} + \frac{3}{2}y^2 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

## 12.3 Independent random variables

Notice that we can always recover the marginal pmf or pdf of a random variable from a joint pmf or pdf. In general we can not go in the opposite direction and construct a joint pmf/pdf from marginal pmf's/pdf's. There is one special case where this can be done, however.

We say two discrete random variables $X$ and $Y$ with joint pmf $p(x,y)$ and respective pmf's $p_X(x)$ and $p_Y(y)$ are ***independent*** if for all choices of $x$ and $y$ we have

$$p(x,y) = p_X(x) \cdot p_Y(y).$$

Similarly, two continuous random variables $X$ and $Y$ with joint pdf $f(x,y)$ and marginal pdf's $f_X(x)$ and $f_Y(y)$ are ***independent*** if for all $x$ and $y$,

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

To motivate this definition, let's think back to what it means to say two events $E$ and $F$ are independent. If $E$ and $F$ are independent, then $\Pr(E \cap F) = \Pr(E) \cdot \Pr(F)$. Notice that for discrete random variables, $p(x,y)$ is the probability $X = x$ and $Y = y$, which we can write as

$$p(x,y) = \Pr(X^{-1}(\{x\}) \cap Y^{-1}(\{y\})),$$

interpreting $X$ and $Y$ as functions defined on a sample space $\Omega$.

The marginal pmf's are exactly equal to

$$p_X(x) = \Pr(X^{-1}(\{x\})), \quad \text{and} \quad p_Y(y) = \Pr(Y^{-1}(\{y\})).$$

Thus $p(x,y) = p_X(x)p_Y(y)$ means

$$\Pr(X^{-1}(\{x\}) \cap Y^{-1}(\{y\})) = \Pr(X^{-1}(\{x\})) \cdot \Pr(Y^{-1}(\{y\}))$$

for all $x$ and $y$. That is, independent random variables correspond to saying the events $X^{-1}(\{x\}), Y^{-1}(\{y\}) \subseteq \Omega$ are indpendent *for all* $x$ and $y$.

(The idea is similar for continuous random variables and pdf's, but slightly obscured by the fact that the pdf doesn't directly tell us probabilities.)

**Example 12.3.**
If $X$ and $Y$ are continuous random variables with joint pdf

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } (x,y) \in [0,1] \times [0,1] \\ 0 & \text{otherwise} \end{cases}$$

are $X$ and $Y$ independent?

First note we can compute the marginal pdf's as

$$f_X(x) = \begin{cases} x + \frac{1}{6} & \text{if } 0 \le x \le 1 \\ 0 & \text{othwerise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{2} + \frac{3y^2}{2} & \text{if } 0 \le y \le 1 \\ 0 & \text{othwerise} \end{cases}$$

For $(x, y) \in [0, 1] \times [0, 1]$, note

$$f_x(x) \cdot f_Y(y) = \frac{x}{2} + \frac{3}{2}xy^2 + \frac{1}{12} + \frac{y^2}{4}$$

which clearly does not equal $f(x, y)$. Thus the random variables are not independent.

**Example 12.4.**
Are the continuous random variables $X$ and $Y$ with joint pdf

$$f(x, y) = \begin{cases} \frac{1}{2}\left(xy^2 - y^2\right) & \text{if } (x, y) \in [1, 3] \times [\sqrt[3]{3/2}, \sqrt[3]{3/2}] \\ 0 & \text{otherwise} \end{cases}$$

independent?

First we must compute the marginal pdf's. For $x \in [1, 3]$ we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x, y)\, dy \\ &= \int_{-\sqrt[3]{3/2}}^{\sqrt[3]{3/2}} \frac{1}{2}\left(xy^2 - y^2\right)\, dy \\ &= \frac{1}{2}\left(\frac{xy^3}{3} - \frac{y^3}{3}\right)\Bigg|_{-\sqrt[3]{3/2}}^{\sqrt[3]{3/2}} \\ &= \frac{1}{2}\left(\frac{x}{2} - \frac{1}{2}\right) - \frac{1}{2}\left(\frac{-x}{2} + \frac{1}{2}\right) \\ &= \frac{x}{2} - \frac{1}{2} \end{aligned}$$

For $y \in [-\sqrt[3]{3/2}, \sqrt[3]{3/2}]$, we have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\ &= \int_1^3 \frac{1}{2}\left(xy^2 - y^2\right)\,dx \\ &= \frac{1}{2}\left(\frac{x^2y^2}{2} - xy^2\right)\,dx \\ &= \frac{1}{2}\left(\frac{9y^2}{2} - 3y^2\right) - \frac{1}{2}\left(\frac{y^2}{2} - y^2\right) \\ &= y^2 \end{aligned}$$

Now notice for $(x,y) \in [1,3] \times [-\sqrt[3]{3/2}, \sqrt[3]{3/2}]$ we have

$$f_X(x)f_Y(y) = \left(\frac{x}{2} - \frac{1}{2}\right)\cdot y^2 = \frac{1}{2}\left(xy^2 - y^2\right) = f(x,y).$$

Of course, for $(x,y)$ outside the rectangle above all the functions are zero and so $f(x,y) = f_X(x)f_Y(y)$ for all $(x,y)$. Thus the random variables are independent.

**Example 12.5.**
Are the discrete random variables with joint pmf

| x \ y | $-1$ | $0$ | $1$ |
|---|---|---|---|
| $1$ | $\frac{1}{12}$ | $\frac{5}{24}$ | $\frac{1}{24}$ |
| $2$ | $\frac{1}{6}$ | $\frac{5}{12}$ | $\frac{1}{12}$ |

independent?

First we compute the marginal pmf's:

$$p_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1 \\ \frac{2}{3} & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} \frac{1}{4} & \text{if } y = -1 \\ \frac{5}{8} & \text{if } y = 0 \\ \frac{1}{8} & \text{if } y = 1 \\ 0 & \text{otherwise} \end{cases}$$

Multiplying $p_X(x) \cdot p_Y(y)$ for all $(x, y)$, we see the random variables are independent.

We can also condition one random variable in terms of another. That is, suppose we have two random variables $X$ and $Y$ and we know the value of $Y$ – say $Y$ has some fixed value $y_0$ – but we don't know the value of $X$. This gives a new random variable which we write as $X|Y = y_0$. What should the pmf (or pdf) of this new random variable be? I.e., how do we compute $\Pr(X \in E|Y = y_0)$.

Notice that if $X$ and $Y$ are both discrete with joint pmf $p(x, y)$ and $Y$ has marginal pmf $p_Y(y)$, then

$$\Pr(X = x|Y = y_0) = \frac{\Pr(X = x \text{ and } Y = y_0)}{\Pr(Y = y_0)} = \frac{\Pr(X^{-1}(\{x\}) \cap Y^{-1}(\{y_0\}))}{\Pr(Y^{-1}(\{y_0\}))} = \frac{p(x, y_0)}{p_Y(y_0)}.$$

This is the pmf of $X|Y = y_0$, which we sometimes denote as $p_{X|Y}(x|y)$:

$$p_{X|Y}(x|y) = \frac{p(x, y)}{p_Y(y)}.$$

For continuous random variables we have a similar definition for the pdf of $X|Y = y$:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

where $f(x, y)$ is the joint pdf of $X$ and $Y$, and $f_Y(y)$ is the marginal pdf of $Y$.

Notice in both cases that $X|Y$ is really a family of random variables: we have one random variable, denoted $X|Y = y$, for each choice of $y$.

**Example 12.6.**
Suppose $X$ and $Y$ are continuous random variables with joint pdf

$$f(x,y) = \begin{cases} x + \frac{3}{2}y^2 & \text{if } (x,y) \in [0,1] \times [0,1] \\ 0 & \text{otherwise} \end{cases}$$

The random variable $X|Y = 0$ is the random variable obtained by fixing $Y = 0$ and allowing $X$ to vary. This has pdf

$$\begin{aligned} f_{X|Y}(x|0) &= \frac{f(x,0)}{f_Y(0)} \\ &= \begin{cases} \frac{x+\frac{3}{2}\cdot 0^2}{\frac{1}{2}+\frac{3}{2}\cdot 0^2} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 2x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

The random variable $X|Y = 1/2$, however, has density

$$\begin{aligned} f_{X|Y}(x|1/2) &= \frac{f(x,1/2)}{f_Y(1/2)} \\ &= \begin{cases} \frac{x+\frac{3}{2}\cdot\left(\frac{1}{2}\right)^2}{\frac{1}{2}+\frac{3}{2}\cdot\left(\frac{1}{2}\right)^2} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{x+\frac{3}{8}}{\frac{1}{2}+\frac{3}{8}} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{8}{7}x + \frac{3}{7} & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Notice that if $X$ and $Y$ happen to be independent, then

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_X(x) \cdot f_Y(y)}{f_Y(y)} = f_X(x).$$

when $X$ and $Y$ are both disjoint. If $X$ and $Y$ are both discrete, then a

similar calculation shows

$$p_{X|Y}(x|y) = p_X(x)$$

when $X$ and $Y$ are independent.

## 12.4 Composition with real-valued functions

Given a function $g : \mathbb{R}^2 \to \mathbb{R}$ and two random variables $X$ and $Y$, the composition $g(X, Y)$ is a new random variable. If $X$ and $Y$ are discrete with joint pmf $p(x, y)$, then $g(X, Y)$ will certainly be discrete and may wonder what its pmf is.

**Theorem 12.1.**
*If $X$ and $Y$ are discrete random variables with joint pmf $p(x, y)$, then for any function $g : \mathbb{R}^2 \to \mathbb{R}$, the composition $g(X, Y)$ is a discrete random variable with pmf*

$$p_g(z) = \sum_{(x,y)\in g^{-1}(\{z\})} p(x, y).$$

*Proof.*
Given $z \in \mathbb{R}$ (we're using $z$ just to prevent confusion with the value of $X$), we need to compute $\Pr(g(X, Y) = z)$. That is, we need to consider the set of $(x, y)$ points which will give $z$ when plugged into $g$; this is precisely $g^{-1}(\{z\})$. Now we simply sum $p(x, y)$ over all points in this preimage. □

**Example 12.7.**
Suppose $g(x, y) = (x^2y + xy^2)^2$, and $X$ and $Y$ are discrete random variables with pmf indicated by the table below.

| $x$ \ $y$ | 1 | 2 | 3 |
|---|---|---|---|
| $-1$ | $1/8$ | $1/4$ | $1/2$ |
| 1 | 0 | $1/8$ | 0 |

What is the pmf $p_g(z)$ of $g(X,Y)$?

Let's first find all the values of $g(X,Y)$ could take on by plugging the $(x,y)$ values of the table above into $g$:

| $x$ \ $y$ | 1 | 2 | 3 |
|---|---|---|---|
| $-1$ | 0 | 4 | 36 |
| 1 | 4 | 36 | 144 |

For each of the four possible values $g(X,Y)$ could take on (0, 4, 36, or 144), we look at what we would $X$ and $Y$ would have to be to get that particular value. The table above basically tells us

$$\begin{aligned} g^{-1}(\{0\}) &= \{(-1,1)\} \\ g^{-1}(\{4\}) &= \{(-1,2),(1,1)\} \\ g^{-1}(\{36\}) &= \{(-1,3),(1,2)\} \\ g^{-1}(\{144\}) &= \{(1,3)\} \end{aligned}$$

For each possibility we sum up the probabilities of these particular $(X,Y)$-values to obtain the pmf

$$p_g(z) = \begin{cases} 1/8 & \text{if } z = 0 \\ 1/4 & \text{if } z = 4 \\ 5/8 & \text{if } z = 36 \\ 0 & \text{if } z = 144 \\ 0 & \text{otherwise} \end{cases}$$

Since $p_g(144) = 0$ we could of course write this in a slightly simpler way as

$$p_g(z) = \begin{cases} 1/8 & \text{if } z = 0 \\ 1/4 & \text{if } z = 4 \\ 5/8 & \text{if } z = 36 \\ 0 & \text{otherwise} \end{cases}$$

We can do something similar for compositions with continuous random variables, but of course there's a bit of calculus involved. First notice that if $X$ and $Y$ are continuous random variables with joint pdf $f(x, y)$, their composition with a function $g : \mathbb{R}^2 \to \mathbb{R}$ may not be continuous. E.g., if we compose with the function

$$g(x, y) = \begin{cases} 1 & \text{if } x \geq y \\ -1 & \text{if } x < y \end{cases}$$

then we certainly will get a discrete random variable. Even if $g(x, y)$ was continuous, the composition $g(X, Y)$ may be discrete. The obvious (boring) example would be if $g(x, y)$ is a constant function. If $g$ is continuous and not constant, however, how would we go about determining the pdf of $g(X, Y)$?

Recalling that the pdf of a continuous random variable is the derivative of the cdf, maybe we should first find the cdf. Let $G$ denote the cdf of $g(X, Y)$:

$$G(z) = \Pr(g(X, Y) \leq z) = \iint_{g^{-1}((-\infty, z])} f(x, y)\, dA.$$

The pdf of $g(X, Y)$, which we'll call $f_g(z)$, is then the derivative of this function:

$$f_g(z) = G'(z) = \frac{d}{dz} \iint_{g^{-1}((-\infty, z])} f(x, y)\, dA.$$

This kind of calculation can be difficult in general, but at least in some particular cases we may be able to compute this pdf.

Calculating pmf's and pdf's of compositions is possible, but rather tedious. Luckily we don't need to do these calculations if all we're interested in is calculating expected values.

**Theorem 12.2.**
*If $X$ and $Y$ are discrete random variables with joint pmf $p(x, y)$ and $g(x, y)$ is any real-valued function, then*

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y) \in \mathbb{R}^2} g(x, y) p(x, y) = \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} g(x, y) p(x, y)$$

*If $X$ and $Y$ are continuous random variables with joint pdf $f(x, y)$ and $g(x, y)$ is any real-valued function such that the composition $g(X, Y)$*

*is a continuous random variable, then*

$$\mathbb{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)f(x,y)\,dA.$$

One important consequence of Theorem 12.2 is the following:

**Corollary 12.3.**
*If $X$ and $Y$ are any two random variables (both discrete or both continuous) on the same sample space $\Omega$, and $\lambda \in \mathbb{R}$ is any constant, then*

$$\mathbb{E}[\lambda X + Y] = \lambda\mathbb{E}[X] + \mathbb{E}[Y].$$

**Remark.**
If you've had linear algebra, all of the remarks above should look like things you've seen before. In particular, multiplying a random variable $X$ by a real number $\lambda$ gives a new random variable $\lambda X$; and adding two random variables $X + Y$ together also gives a random variable. That is, the set of all random variables on a given sample space $\Omega$ is a real vector space. Furthermore, Corollary 12.3 says that expectation, $\mathbb{E}$, is a linear transformation from this vector space to the vector space $\mathbb{R}$.

## 12.5 Covariance and correlation

We now introduce a number associated to each pair of random variables $X$ and $Y$ which gives us a measure of how changes in one random variable relate to changes in the other. This number is called the *covariance* of $X$ and $Y$.

To be more precise, the ***covariance*** of $X$ and $Y$, denoted $\mathrm{Cov}(X,Y)$, is defined to be the expected value of $(X-\mu_X)\cdot(Y-\mu_Y)$ where $\mu_X = \mathbb{E}[X]$

and $\mu_Y = \mathbb{E}[Y]$,

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X) \cdot (Y - \mu_Y)].$$

Before mentioning some basic properties of this quantity, let's try to get some intuition for what the covariance actually measures. Notice that for any random output of our random variable $X$, the factor $X - \mu_X$ is positive if $X$ is greater than its mean and negative if it is smaller than its mean. Similarly for $Y - \mu_Y$. The product $(X - \mu_X) \cdot (Y - \mu_Y)$ is positive if $X - \mu_X$ and $Y - \mu_Y$ have the same sign (both positive or both negative), and negative if they have different signs. Just paying attention to the sign of these values and ignoring their magnitude, $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ tells us if the signs agree (in which case this expected value is positive) or disagree (so the expected value is negative) on average.

We'll see some examples of this in just a minute which should help give some intuition for covariance, but first we mention a couple of simple properties that will make our calculations slightly easier.

Let's notice that the covariance of $X$ and $X$ (i.e., plugging in $X$ for $Y$ as well) gives the variance of $X$:

$$\begin{aligned}\mathrm{Cov}(X, X) &= \mathbb{E}[(X - \mu_X) \cdot (X - \mu_X)] \\ &= \mathbb{E}[(X - \mu_X)^2] \\ &= \mathrm{Var}(X).\end{aligned}$$

Just as there's a minor shortcut for computing variance, there's a corresponding shortcut for computing covariance.

**Proposition 12.4.**
*For any two random variables $X$ and $Y$,*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

*Proof.*

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\ &= \mathbb{E}[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - \mu_X\mathbb{E}[Y] + \mu_X\mu_Y \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

□

Before going any further, let's consider a couple of very simple examples.

**Example 12.8.**
Suppose $X$ and $Y$ are discrete random variables whose joint pmf is the following:

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(1,1),(2,2),(3,3),(4,4)\} \\ 0 & \text{otherwise} \end{cases}$$

To compute the covariance we need the expected value of each of $X$ and $Y$ which requires that we compute their marginals. It is easy to check in this case the marginals are

$$p_X(x) = \begin{cases} 1/4 & \text{if } x \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 1/4 & \text{if } y \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases}$$

Now we compute the expected values to obtain

$$\mathbb{E}[X] = \mathbb{E}[Y] = \frac{5}{2}.$$

and the expected value of $XY$ is

$$\mathbb{E}[XY] = \frac{1+4+9+16}{4} = \frac{30}{4} = \frac{15}{2}.$$

So the covariance here is $\frac{15}{2} - \frac{5}{2} \cdot \frac{5}{2} = \frac{5}{4}$.

Notice in this example that as the $X$ values increased the $Y$ values increased in the same way and we had a positive covariance. What happens if instead the $Y$ values decrease as $X$ decreases?

Suppose now that $X$ and $Y$ have joint pmf

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(1,4),(2,3),(3,2),(4,1)\} \\ 0 & \text{otherwise} \end{cases}$$

Now the marginals are still

$$p_X(x) = \begin{cases} 1/4 & \text{if } x \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 1/4 & \text{if } y \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases}$$

and so the expected value of $X$ and $Y$ are still $\mathbb{E}[X] = \mathbb{E}[Y] = 5/2$. The expected value of $XY$, however, is

$$\mathbb{E}[XY] = \frac{4+6+6+4}{4} = 5.$$

The covariance is now $5 - \frac{25}{4} = \frac{-5}{4}$.

Notice in the examples above that the covariance was positive when the $X$ and $Y$ values increased together, but was negative when the $Y$-values decreased as $X$ increased.

That is, covariance (or at least the sign of covariance), tells us if $X$ and $Y$ increase together (equivalently, decrease together) or if one increases while the other decreases.

**Exercise 12.1.**
Compute the covariance of discrete random variables $X$ and $Y$ whose joint pmf is

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(1,1),(2,4),(3,9),(4,16)\} \\ 0 & \text{otherwise} \end{cases}$$

One important property of covariance is that it is always zero for independent random variables:

**Proposition 12.5.**
*If $X$ and $Y$ independent, then* $\mathrm{Cov}(X, Y) = 0$.

*Proof.*
We will prove this for the case when $X$ and $Y$ are continuous; the discrete case is very similar. Let $f(x, y)$ be the joint pdf of $X$ and $Y$, and $f_X(x)$ and $f_Y(y)$ the marginal pdf's. Suppose $X$ and $Y$ are independent so $f(x, y) = f_X(x)f_Y(y)$, and we simply compute

$$\begin{aligned}
\mathbb{E}[XY] &= \iint_{\mathbb{R}^2} xyf(x, y)\, dA \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} xf_X(x)\, dx \cdot \int_{-\infty}^{\infty} yf_Y(y)\, dy \\
&= \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

Now applying our formula for covariance above we have

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

□

**Example 12.9.**

Suppose $X$ and $Y$ are continuous random variables with joint pdf

$$f(x,y) = \begin{cases} 3x & \text{if } 0 \le y \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the covariance of $X$ and $Y$?

First let's notice we must compute $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, which requires that we find the marginal pdf's.

For $X$ we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\,dy \\ &= \begin{cases} \int_0^x 3x\,dy & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 3xy\Big|_0^x & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} 3x^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and so

$$\begin{aligned} \mathbb{E}[X] &= \int_0^1 x 3x^2\,dx \\ &= \int_0^1 3x^3\,dx \\ &= \frac{3x^4}{4}\Big|_0^1 \\ &= \frac{3}{4} \end{aligned}$$

For $Y$,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,dx \\ &= \begin{cases} \int_y^1 3x dx & \text{if } 0 \le y1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \left.\frac{3x^2}{2}\right|_y^1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{3}{2}\left(1-y^2\right) & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}[Y] &= \int_0^1 y\frac{3}{2}\left(1-y^2\right)\,dy \\ &= \int_0^1 \frac{3}{2}\left(y-y^3\right)\,dy \\ &= \frac{3}{2}\left(\frac{y^2}{2}-\frac{y^4}{4}\right)\Bigg|_0^1 \\ &= \frac{3}{2}\left(\frac{1}{2}-\frac{1}{4}\right) \\ &= \frac{3}{8} \end{aligned}$$

Now to compute the covariance we must also compute $\mathbb{E}[XY]$:

$$\mathbb{E}[XY] = \iint_{\mathbb{R}^2} xyf(x,y)\,dA$$

Now we compute the covariance as follows,

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \frac{3}{10} - \frac{3}{4}\cdot\frac{3}{8} \\ &= \frac{3}{10} - \frac{9}{32} \\ &= \frac{96-90}{320} \\ &= \frac{6}{320} \\ &= \frac{3}{160}\end{aligned}$$

We noted above that if $X$ and $Y$ are independent, then their covariance must be zero. The next example illustrates that the converse *is not* true: $X$ and $Y$ may have zero covariance without being independent.

**Example 12.10.**
Suppose $X$ and $Y$ are two discrete random variables with joint pmf

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(0,0),(1,1),(1,-1),(2,0)\} \\ 0 & \text{otherwise} \end{cases}$$

Are $X$ and $Y$ independent? If $X$ and $Y$ are not independent, what is their covariance?

To determine if $X$ and $Y$ are independent or not we must compute their marginal pmf's:

$$\begin{aligned}f_X(x) &= \sum_{y\in\mathbb{R}} f(x,y) \\ &= f(x,0) + f(x,1) + f(x,-1) \\ &= \begin{cases} 1/4 & \text{if } x = 0 \\ 1/2 & \text{if } x = 1 \\ 1/4 & \text{if } x = 2 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

$$\begin{aligned} f_Y(y) &= \sum_{x \in \mathbb{R}} f(x,y) \\ &= f(0,y) + f(1,y) + f(2,y) \\ &= \begin{cases} 1/4 & \text{if } x = -1 \\ 1/2 & \text{if } x = 0 \\ 1/4 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Now we can easily see that $f(x,y) \neq f_X(x) f_Y(y)$. For example, $f(0,0) = 1/4$, while $f_X(0) f_Y(0) = 1/4 \cdot 1/2 = 1/8$. So, $X$ and $Y$ are not independent.

For the covariance we simply compute the necessary expectations:

$$\begin{aligned} \mathbb{E}[X] &= 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1 \\ \mathbb{E}[Y] &= -1 \cdot 1/4 + 0 \cdot 1/2 + 1 \cdot 1/4 = 0 \\ \mathbb{E}[XY] &= 0 \cdot 0 \cdot 1/4 + 1 \cdot 1 \cdot 1/4 + 1 \cdot (-1) \cdot 1/4 + 2 \cdot 0 \cdot 1/4 \\ &= 0 \end{aligned}$$

Now we compute the covariance as

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0 - 1 \cdot 0 = 0.$$

So $\text{Cov}(X,Y) = 0$, even though $X$ and $Y$ are not independent.

The following theorem tells us that we can break up covariance calculations when we have a function of random variables by splitting up sums and differences, and pulling out scalars.

**Theorem 12.6.**
*If $X$, $Y$, and $Z$ are any three random variables defined on the same sample space, and if $\lambda \in \mathbb{R}$ is any real number, then*

1. $\text{Cov}(X,Y) = \text{Cov}(Y,X)$
2. $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$

3. $\text{Cov}(\lambda X, Y) = \lambda \text{Cov}(X, Y)$

**Exercise 12.2.**
Prove Theorem 12.6.

**Remark.**
If you've taken linear algebra before, the properties in Theorem 12.6 should look familiar: these are exactly the axioms for an inner product on a real vector space. That is, covariance is an inner product on the vector space of all random variables on a given sample space. This has some very important consequences which we won't have time to discuss in this class, but it essentially means there's a geometric way of thinking of the space of random variables. (Inner products supply us with a notion of angle and length.) One consequence, for example, is that indpenendent random variables are orthogonal!

We saw earlier that the sign of covariance told us something about how two random variables $X$ and $Y$ increase/decrease together. The magnitude of the covariance is less important, though, and so we now introduce a sort of "normalization" of the covariance which lets us ignore the magnitude.

The ***correlation coefficient*** (or simply ***correlation***) between two random variables $X$ and $Y$, denoted $\text{Corr}(X, Y)$, is defined to be the quantity

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X$ and $\sigma_Y$ are the standard deviation of $X$ and $Y$, respectively.

**Example 12.11.**
Suppose $X$ and $Y$ are discrete random variables with joint pmf as indicated below:

| $x$ \ $y$ | 1 | 2 | 3 |
|---|---|---|---|
| $-1$ | $1/8$ | $1/4$ | $1/2$ |
| 1 | 0 | $1/8$ | 0 |

What is their correlation?

To find the correlation we must find the covariance, which requires us to find the expected value of each random variable, and the standard deviation. To do these calculations we must calculate the marginal pmf's of $X$ and $Y$:

$$p_X(x) = \begin{cases} 7/8 & \text{if } x = -1 \\ 1/8 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 1/8 & \text{if } y = 1 \\ 3/8 & \text{if } y = 2 \\ 1/2 & \text{if } y = 3 \\ 0 & \text{otherwise} \end{cases}$$

Now we find the expected value of $X$ and $Y$:

$$\mathbb{E}[X] = -1 \cdot \frac{7}{8} + 1 \cdot \frac{1}{8} = \frac{-3}{4}$$
$$\mathbb{E}[Y] = 1 \cdot \frac{1}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{2} = \frac{19}{8}$$

Now we also need to find the variance of $X$ and $Y$ so we can find the

standard deviations:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \left((-1)^2 \cdot \frac{7}{8} + 1^2 \cdot \frac{1}{8}\right) - \left(\frac{-3}{4}\right)^2 \\ &= 1 - \frac{9}{16} \\ &= \frac{7}{16}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \left(1^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{2}\right) - \left(\frac{19}{8}\right)^2 \\ &= \frac{49}{8} - \frac{361}{64} \\ &= \frac{2775}{64}\end{aligned}$$

The standard deviations are thus

$$\sigma_X = \frac{\sqrt{7}}{4} \qquad \text{and} \qquad \sigma_Y = \frac{\sqrt{2775}}{8}.$$

We also need $\mathbb{E}[XY]$ to compute the covariance:

$$\begin{aligned}\mathbb{E}[XY] &= \sum_{(x,y)\in\mathbb{R}^2} xyp(x,y) \\ &= (-1)\cdot\frac{1}{8} + (-2)\cdot\frac{1}{4} + (-3)\cdot\frac{1}{2} + 2\cdot\frac{1}{8} \\ &= -1/8 - 4/8 - 12/8 + 2/8 \\ &= \frac{-15}{8}.\end{aligned}$$

The covariance is thus

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \frac{-15}{8} - \frac{-3}{4}\cdot\frac{19}{8} && = \frac{-15}{8} + \frac{57}{32} \\ &= \frac{-60+57}{32} \\ &= \frac{-3}{32}.\end{aligned}$$

Now we have all the pieces to compute the correlation:

$$\begin{aligned}\operatorname{Corr}(X,Y) &= \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{-3/32}{\sqrt{7}/4 \cdot \sqrt{2775}/8} \\ &= \frac{-3/32}{\sqrt{19425}/32} \\ &= \frac{-3}{\sqrt{19425}}\end{aligned}$$

Computationally, all of the hard work comes from computing the covariance; we then just divide this covariance by some particular (which also requires a bit of work). The basic properties of covariance thus carry over to correlation, but we also have a few new properties:

**Theorem 12.7.**
*Let $X$ and $Y$ be two random variables on the same sample space $\Omega$. Then we have the following four properties:*

1. $-1 \leq \operatorname{Corr}(X,Y) \leq 1$
2. *For any constants $a, b, c, d \in \mathbb{R}$, we have* $\operatorname{Corr}(aX+b, cY+d) = \operatorname{Corr}(X,Y)$.
3. *If $X$ and $Y$ are independent, then* $\operatorname{Corr}(X,Y) = 0$.

4. $\mathrm{Corr}(X, Y) = \pm 1$ *if and only if* $Y = aX + b$ *for some constants* $a$ *and* $b$.

These properties are a little bit trickier to prove from first principles (i.e., without appealing to some slightly more advanced math), so we'll skip the proof but make one little observation about how to prove the first part of Theorem 12.7 in the remark below.

**Remark.**
The first part of Theorem 12.7 is actually the Cauchy-Schwarz inequality for inner products. In particular, since Cov is an inner product, $\sqrt{\mathrm{Cov}(X, X)} = \sqrt{\mathrm{Var}(X)} = \sigma_X$ is a norm. The Cauchy-Schwarz inequality says that for any vectors $v, w$ in an inner product space,

$$|\langle v, w \rangle| \leq \|v\| \, \|w\|$$

In our particular setting where the inner product is given by covariance and the norm is the standard deviation, this becomes

$$|\mathrm{Cov}(X, Y)| \leq \sigma_X \sigma_Y.$$

We can rewrite this as

$$-\sigma_X \sigma_Y \leq \mathrm{Cov}(X, Y) \leq \sigma_X \sigma_Y.$$

Dividing through by $\sigma_X \sigma_Y$ gives the desired inequality:

$$-1 \leq \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \mathrm{Corr}(X, Y) \leq 1.$$

Properties (1) and (4) from Theorem 12.7 tell us that correlation is a measure of the linear relationship between $X$ and $Y$. If $|\mathrm{Corr}(X, Y)|$ is very close to one, then $X$ and $Y$ are very close to being related by a linear function; the sign tells us whether $X$ and $Y$ are positively or negatively proportional.

**Exercise 12.3.**
Compute the correlation coefficients for the discrete random variables $X$ and $Y$ with joint pmf

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(1,1),(2,2),(3,3),(4,4)\} \\ 0 & \text{otherwise} \end{cases}$$

Repeat the exercise when the joint pmf is

$$p(x,y) = \begin{cases} 1/4 & \text{if } (x,y) \in \{(1,4),(2,3),(3,2),(4,1)\} \\ 0 & \text{otherwise} \end{cases}$$

## 12.6 Linear combinations of random variables

There are a few other important properties of covariance we need to be aware of, but in order to state them concisely we need to give a definition.

A ***linear combination*** of a finite number of random variables $X_1$, $X_2$, ..., $X_n$ is simply a sum where each random variable $X_i$ is multiplied by some constant $\lambda_i$ (we allow different $\lambda_i$'s for different $X_i$'s), then these are all added together. For example,

$$2X_1 - X_2 + 7X_3$$

is a linear combination, and so is

$$\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

(This second linear combination will turn out to be very important later.)

It follows from our earlier work that expected values split up nicely for linear combinations.

**Lemma 12.8.**
*For any collection of $n$ random variables $X_1$, $X_2$, ..., $X_n$ and any*

*collection of $n$ real numbers $\lambda_1$, $\lambda_2$, ..., $\lambda_n$ we have*

$$\mathbb{E}[\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n] = \lambda_1 \mathbb{E}[X_1] + \lambda_2 \mathbb{E}[X_2] + \cdots + \lambda_n \mathbb{E}[X_n]$$

*Proof.*
We will prove this only in the case when the $X_i$ are discrete; the proof when they are continuous is similar.

Suppose $p(x_1, ..., x_n)$ is the joint pmf of the random variables. For notational convenience, let's write $\vec{x}$ for $(x_1, x_2, ..., x_n)$. The expected value above then becomes

$$\mathbb{E}[\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n] = \sum_{\vec{x} \in \mathbb{R}^n} (\lambda_1 x_1 + \cdots + \lambda_n x_n) p(\vec{x}).$$

Now we split this sum up and factor out the $\lambda_i$'s to write the expected value as

$$\lambda_1 \sum_{\vec{x} \in \mathbb{R}^n} x_1 p(\vec{x}) + \lambda_2 \sum_{\vec{x} \in \mathbb{R}^n} x_2 p(\vec{x}) + \cdots + \lambda_n \sum_{\vec{x} \in \mathbb{R}^n} x_n p(\vec{x}).$$

We claim each of these sums is in fact $\mathbb{E}[X_i]$. To see this, note we could rewrite the sum as

$$\sum_{\vec{x} \in \mathbb{R}^n} x_i p(\vec{x}) = \sum_{x_1 \in \mathbb{R}} \sum_{x_2 \in \mathbb{R}} \cdots \sum_{x_n \in \mathbb{R}} x_i p(x_1, x_2, ..., x_n).$$

Rearrange this sum so the $x_i$ terms come first, and factor $x_i$ out of all of the remaining sums to obtain

$$\sum_{x_i \in \mathbb{R}} x_i \left( \sum_{x_1 \in \mathbb{R}} \sum_{x_2 \in \mathbb{R}} \cdots \sum_{x_{i-1} \in \mathbb{R}} \sum_{xi+1 \in \mathbb{R}} \cdots \sum_{x_n \in \mathbb{R}} p(x_1, ..., x_n) \right).$$

Notice the sum on the right is precisely the marginal pmf of $X_i$, so the expression above is simply

$$\sum_{x_i \in \mathbb{R}} x_i p_{X_i}(x_i) = \mathbb{E}[X_i].$$

Plugging this into the expression above proves the lemma. □

So, expected values break up nicely for a linear combination of random variables. There is a similar way to break up variances, but it's not quite as simple.

**Lemma 12.9.**
*For any collection of $n$ random variables $X_1$, $X_2$, ..., $X_n$ and any collection of $n$ real numbers $\lambda_1$, $\lambda_2$, ..., $\lambda_n$ we have*

$$\text{Var}\left(\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \text{Cov}(X_i, X_j)$$

*Proof.*
We essentially just apply properties of covariance from Theorem 12.6

together with the fact $\text{Var}(X) = \text{Cov}(X, X)$ to verify the claim.

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^{n} \lambda_i X_i\right) &= \text{Cov}\left(\sum_{i=1}^{n} \lambda_i X_i, \sum_{j=1}^{n} \lambda_j X_j\right) \\
&= \sum_{i=1}^{n} \text{Cov}\left(\lambda_i X_i, \sum_{j=1}^{n} \lambda_j X_j\right) \\
&= \sum_{i=1}^{n} \lambda_i \text{Cov}\left(X_i, \sum_{j=1}^{n} \lambda_j X_j\right) \\
&= \sum_{i=1}^{n} \lambda_i \text{Cov}\left(\sum_{j=1}^{n} \lambda_j X_j, X_i\right) \\
&= \sum_{i=1}^{n} \lambda_i \cdot \left(\sum_{j=1}^{n} \text{Cov}\left(\lambda_j X_j, X_i\right)\right) \\
&= \sum_{i=1}^{n} \lambda_i \cdot \left(\sum_{j=1}^{n} \lambda_j \text{Cov}\left(X_j, X_i\right)\right) \\
&= \sum_{i=1}^{n} \lambda_i \cdot \left(\sum_{j=1}^{n} \lambda_j \text{Cov}\left(X_i, X_j\right)\right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \text{Cov}\left(X_i, X_j\right)
\end{aligned}$$

□

An important corollary is the following.

**Corollary 12.10.**
*If $X_1$, $X_2$, ..., $X_n$ are mutually independent, then for any real numbers $\lambda_1$, $\lambda_2$, ..., $\lambda_n$,*

$$\text{Var}\left(\sum_{i=1}^{n} \lambda_i X_i\right) = \sum_{i=1}^{n} \lambda_i^2 \text{Var}(X_i)$$

**Exercise 12.4.**
Prove Corollary 12.10.

## 12.7 The strong law of large numbers and central limit theorem

We now quickly mention two theorems that are arguably the most important in all of statistics. Most of the theorems we have seen in class are basic properties of the constructions we have introduced, but the next two theorems are much deeper. These two theorems are the real work horses for statistics; most of statistics would not be possible without these theorems. We will, unfortunately, need to treat these theorems as blackboxes, however: their proofs are too advanced for this class, though the results are fundamental.

In this section we will just state the theorems, since they naturally fit in with the material about jointly distributed random variables, though we won't see any significant examples until the next chapter. So, while you'll need to take it on faith these theorems are interesting and important for the moment, in the next chapter we will start using these theorems to study statistics.

Before we can state these theorems, we need one more definition. We say a sequence of random variables $X_1$, $X_2$, $X_3$, ... are ***independent and identically distributed***, often abbreviated ***IID***, if the $X_i$ are all mutually independent and they have the same distribution (i.e., they're all discrete or all continuous; if they are discrete, they all have the same pmf; if continuous, they all have the same pdf).

The strong law of large numbers says that if we have such an IID collection of random variables, then their average approaches the expected value of any one (and hence all) of the $X_i$.

**Theorem 12.11** (The strong law of large numbers)**.**
*Let $X_1$, $X_2$, $X_3$, ... be a sequence of IID random variables, and let $\mu$*

*denote the expected value of any of the $X_i$. Then, with probability one,*

$$\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mu.$$

Let's first notice that the statement of the theorem is quantified by the phrase *with probability one*. We have a random collection of numbers, and it could happen that by sheer bad luck we actually got values for $X_1$, $X_2$, ... so that the limit above was not $\mu$. For example, suppose each $X_i$ was a binomial random variable with parameter $1/2$. That is, $\Pr(X_i = 1) = 1/2$ and $\Pr(X_i = 0) = 1/2$ for each $i$. The strong law of large numbers tells us that if we select longer and longer sequences of numbers, each of which is one of these binomial random variables, then when we average these numbers together we should get the mean of that binomial, $1/2$.

It is conceivable, however, that we could select $X_1$, $X_2$, $X_3$, ... so that each equalled 0, and so the limit is zero instead of $1/2$. However, the probability this happens is zero. In this particular case this is easy to see: as the $X_i$ are all mutually independent, we see

$$\begin{aligned}
&\Pr(X_1 = 0 \text{ and } X_2 = 0 \text{ and } X_3 = 0 \text{ and } \cdots X_n = 0) \\
&= \Pr(X_1 = 0) \cdot \Pr(X_2 = 0) \cdot \Pr(X_3 = 0) \cdots \Pr(X_n = 0) \\
&= \frac{1}{2^n}.
\end{aligned}$$

As $n$ goes to infinity, this goes to zero. In fact, for any particular sequence of zeros and ones, the probability we get that particular sequence is zero by the same argument. The strong law of large numbers is saying something more, though: it is saying that when you pick these random numbers, with probability one you will pick numbers whose average goes the value $\mu$. It's not guaranteed this will happen, but the probability it doesn't happen is staggeringly small – so small the probability we don't average out to $\mu$ is zero. This is what the *with probability one* part of the strong law of large numbers says. (Sometimes this is also called ***convergence almost surely***: the strong law of large numbers says the limit above "almost surely" converges to $\mu$.)

The take-away from the strong law of large numbers is that if we have a sequence of random variables which we assume are IID, but we don't know what the pmf/pdf is and can't compute $\mu$, then we can estimate $\mu$ by averaging several $X_i$ values together.

Notice the quantity appearing in the strong law of large numbers,

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

is itself a random variable. The next theorem tells us something more precise about the distribution of these random variables for large values of $n$.

**Theorem 12.12** (The central limit theorem)**.**
*Let $X_1$, $X_2$, $X_3$, ... be a sequence of IID random variables. Let $\mu = \mathbb{E}[X_i]$ be the common mean of these random variables, and $\sigma = \sqrt{\mathrm{Var}(X_i)}$ the common standard deviation.*

*For each integer $n \geq 1$, define a new random variable $Z_n$ as*

$$Z_n = \sqrt{n} \cdot \frac{X_1 + X_2 + \cdots + X_n - n\mu}{n\sigma},$$

*and let $F_n$ denote the cdf of this random variable. Then for every $x \in \mathbb{R}$,*

$$\lim_{n \to \infty} F_n(x) = \Phi(x)$$

*where $\Phi$ is the cdf of the standard normal, $Z$.*

The central limit theorem takes a little bit of work to decipher, but once you understand what it says, it almost seems like magic. The theorem says that if we take *any* sequence of IID random variables (literally any sequence of weird random variables you like, discrete, continuous, whatever) and then perform a certain normalization to that sequence, your normalized random variables approach the standard normal. This is incredible because it means we can actually use knowledge of the standard normal to help us understand *any* sequence of random variables.

We should point out that there are actually several different versions of the central limit theorem, and if you looked in four different textbooks you might see four slightly different variations on the theorem above. The version above is the most correct and precise version of the theorem we can give right now, but there's another version that is sometimes easier to think about.

**Theorem 12.13** (Alternative central limit theorem)**.**
*Let $X_1$, $X_2$, $X_3$, ... be a sequence of IID random variables. Then for "sufficiently large" values of $n$, the random variable*

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

*is approximately the normal random variable $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.*
*(Here, $\mu = \mathbb{E}[X_i]$ and $\sigma = \sqrt{\mathrm{Var}(X_i)}$.)*

This alternative version of the central limit theorem tells us that if $n$ is large enough, the random variable $\frac{1}{n}(X_1 + \cdots + X_n)$ is essentially a normal random variable. This is helpful because if we want to know the probability this random variable takes on a value in a certain range, the central limit theorem says we can estimate that probability by computing the probability for the corresponding normal random variable. (Of course, we do that by standardizing the normal. The subtraction of $n\mu$, division by $n\sigma$ and multiplication by $\sqrt{n}$ in the earlier version of the central limit theorem simply has that standardization worked out.)

**Remark.**
In some basic statistics courses you are told the approximation from the alternative version of the central limit theorem applies as long as $n$ is at least thirty. This is just a heuristic and not a mathematical fact. That is, for many "real world" random variables $n \geq 30$ seems to be large enough that the normal is a reasonable approximation for $\frac{1}{n}(X_1+X_2+\cdots+X_n)$. However, it is possible to have random variables where $n$ might be one-hundred billion before $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is well approximated by the normal. The "sufficiently large" part of our alternative central limit theorem is sweeping this under the rug. There does exist some large enough value of $n$ so that $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is as close to a normal random variable as you'd like, although that large enough $n$ depends on exactly what the distribution of the $X_i$ is.

## 12.8 Practice problems

**Problem 12.1.**
Suppose $X$ and $Y$ are two discrete random variables whose joint probability density function, $p(x, y) = P(X = x \text{ and } Y = y)$, is given by the following table:

| $x$ \ $y$ | $-2$ | $-1$ | $1$ | $2$ |
|---|---|---|---|---|
| 0 | 0.1 | 0 | 0.05 | 0.05 |
| 1 | 0 | 0.1 | 0.1 | 0 |
| 2 | 0.07 | 0.02 | 0.04 | 0.07 |
| 3 | 0.1 | 0.05 | 0.03 | 0.02 |

(a) Determine the probability that $X \leq 2$ and $Y < 1$.

(b) Compute the marginal pdf of $X$.

(c) Compute the marginal pdf of $Y$.

(d) Compute $\mathbb{E}[XY]$.

(e) Compute the pdf of the random variable $X + Y$.

**Problem 12.2.**
Suppose $X$ and $Y$ are continuous random variables with joint pdf

$$f(x, y) = \begin{cases} kxy & \text{if } 0 \leq x,\ 0 \leq y, \text{ and } x + y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

for some constant $k$.

(a) Determine which value of $k$ makes the $f(x, y)$ above a pdf.

(b) Compute $P(Y \geq X)$.

(c) Are $X$ and $Y$ independent?

**Problem 12.3.**
Suppose a point in the unit disc in the plane,

$$\{(x, y) \in \mathbb{R}^2 \,\big|\, 0 \leq x^2 + y^2 \leq 1\},$$

is randomly selected where the joint pdf of the $x$ and $y$ coordinates of the point is given by

$$f(x, y) = \frac{1}{\pi}.$$

For each $0 \leq r \leq 1$, find the probability the randomly selected point is distance at most $r$ from the origin.

**Problem 12.4.**
Suppose $X$ and $Y$ are two independent Poisson random variables with parameters $\lambda_X$ and $\lambda_Y$. Determine the pdf of $X + Y$.

(Hint: Note if $X + Y = n$, then for some $0 \leq k \leq n$ we have $X = k$ and $Y = n - k$.)

# Part IV

# Statistics

# 13 Introduction to Statistics

*The beauty of mathematics only shows itself to more patient followers.*

Maryam Mirzakhani
*Interview in Clay Mathematics Institute Annual Report, 2008*

## 13.1 The idea of statistics

Statistics is the study of data, and in particular the study of how to deduce information from data. This data is obtained from a ***population***, which is some collection of all of the "objects" we care about. For example, all cars driven in the state of Indiana; all ears of corn in a farm; or all of the college football players in the United States.

In an ideal world, we might like to have total information about all members of a population. For example, we might want to know the tread level of the tires on every car in Indiana; or the sugar content of every ear of corn in a field; or the fractional anisotropy value[1] of each football player in the country. Having this kind of complete information is called a ***census***, but it is often impossible (or at least very impractical) to perform a census on a given population. E.g., it is not realistic for a farmer to actually measure the sugar content of every ear of corn in their field.

Instead of performing a census, we instead consider a more manageable collection of values obtained from a subset of the population called a ***sample***. The goal of statistics is to take information obtained from a sample and use it to deduce information about the population as a whole.

Broadly, there are two basic types of statistics:

- ***Descriptive statistics*** is concerned with summarizing data. This may be done graphically (e.g., with histograms, boxplots, pie charts, etc.) or numerically (e.g., computation of a mean, median, mode, and standard deviation). This sort of descriptive statistics is likely what you've seen if you've taken any kind of statistics class before in high-school or college.

[1]The fractional anisotropy is a number which indicates damage to white matter in the brain, and requires an MRI to determine. This value might be relevant to researchers studying the cause of chronic truamatic encephalopathy, a type of brain disease common among people who have repeated concussions, such as football players.

- ***Inferential statistics*** uses data from a sample to infer information about the population. For example, if there is a particular parameter of the population we are interested in, we may be able to give a range of possible values for that parameter based on data obtained in a sample. (This is called a ***confidence interval*** and one of the topics we will discuss in detail soon.)

Descriptive statistics is relatively elementary material, and so we are not going to spend any time discussing it, and will instead jump straight into inferential statistics. It's worth point out, though, that if you took a class in descriptive statistics before and found it rather uninteresting (e.g., just memorizing lots of formulas or learning how to get a computer or calculator to generate charts or compute medians), then you may find inferential statistics much more interesting.

Whereas descriptive statistics is basically an exercise in memorizing definitons and formulas, inferential statistics is serious mathematics. All of the probability theory and random variables we have discussed up to this point are just the background material needed to make inferential statistics precise.

In particular, when we take a sample from a population and then measure some quantity for the members of our sample, we treat these values are random variables. For instance, if we measure the sugar content of ears of corn in a sample of 100 ears from a field, we don't know what the sugar content will be until we actually measure it and for this reason think of it as a random variable. We also assume that our sample is representative of the population as a whole, and so assume these random variables are IID. The goal will be to determine the distribution of the underlying random variables based on the data we obtain from those IID random variables, and this is where the strong law of large numbers and central limit theorem are useful.

## 13.2 What is a statistic?

Before going any further, it's worth taking the time to point out a potential for confusion. The word "statistics" has two meanings: one is a general study of data, but the other is a precise mathematical definition.

A ***statistic*** is a function (or sometimes the output of such a function) that depends *only* on the values determined by a sample. It's important to realize this really depends *only* on sample values and not unknown parameters of the population! That is, we have some collection of sample data

(aka, random variables), say $X_1$, $X_2$, ..., $X_n$, and then we have some function of these values, $g(X_1, X_2, ..., X_n)$. This function $g$, or the output of the function after we plug in our sample data, is a statistic. Notice that since a statistic is a function of random variables, it is itself a random variable, and so we can ask questions about its expected value, variance, whether it's discrete or continuous, what its pmf or pdf is, etc.

We have already alluded to one particular example of a statistic earlier, though we didn't call it a statistic at that time. The ***sample mean*** of $X_1$, $X_2$, ..., $X_n$ is the statistic

$$\frac{X_1 + X_2 + \cdots + X_n}{n}$$

and is often denoted $\overline{X}$, where the number of samples, $n$, is often understood from context.

Let's have a simple example of this particular statistic and determine its sampling distribution to help make all of this more concrete.

**Example 13.1.**
Suppose that the students at a particular university have the following distribution of GPA's:

- 5% of students have a GPA of 1.
- 30% of students have a GPA of 2.
- 60% of students have a GPA of 3.
- 5% of students have a GPA of 4.

Now suppose we randomly select two students from this university at random and record their GPA's; let $X_1$ be the GPA of the first student, and $X_2$ the GPA of the second student. The statistic $\overline{X}$ is the average of these two GPA's, whatever they happen to be,

$$\overline{X} = \frac{X_1 + X_2}{2}.$$

Notice this is a random variable: the values $X_1$ and $X_2$ are random (we don't know beforehand which students will select or what their GPA will be), and so $\overline{X}$ is also random. However, there are only so many choices for what $X_1$ and $X_2$ could be, and we know the probability of

each of these possibilities, so we can work out the probability for each possibility of $\overline{X}$.

In the table below we look at all possibilities of $X_1$ and $X_2$, compute the corresponding $\overline{X}$, and then compute the probability of this particular choice of $X_1$ and $X_2$. (Recall we are assuming now that $X_1$ and $X_2$ are IID, so we can just multiply the probability $X_1$ equals a given value and the probability $X_2$ equals a given value.)

| $X_1$ | $X_2$ | $\overline{X} = \frac{X_1+X_2}{2}$ | Probability |
|---|---|---|---|
| 1 | 1 | 1 | $0.5 \cdot 0.5 = 0.0025$ |
| 1 | 2 | 1.5 | $0.5 \cdot 0.3 = 0.015$ |
| 1 | 3 | 2 | $0.5 \cdot 0.6 = 0.03$ |
| 1 | 4 | 2.5 | $0.5 \cdot 0.5 = 0.0025$ |
| 2 | 1 | 1.5 | $0.3 \cdot 0.05 = 0.015$ |
| 2 | 2 | 2 | $0.3 \cdot 0.3 = 0.09$ |
| 2 | 3 | 2.5 | $0.3 \cdot 0.6 = 0.18$ |
| 2 | 4 | 3 | $0.3 \cdot 0.05 = 0.015$ |
| 3 | 1 | 2 | $0.6 \cdot 0.05 = 0.03$ |
| 3 | 2 | 2.5 | $0.6 \cdot 0.3 = 0.18$ |
| 3 | 3 | 3 | $0.6 \cdot 0.6 = 0.36$ |
| 3 | 4 | 3.5 | $0.6 \cdot 0.05 = 0.03$ |
| 4 | 1 | 2.5 | $0.05 \cdot 0.05 = 0.0025$ |
| 4 | 2 | 3 | $0.05 \cdot 0.3 = 0.015$ |
| 4 | 3 | 3.5 | $0.05 \cdot 0.6 = 0.03$ |
| 4 | 4 | 4 | $0.05 \cdot 0.05 = 0.0025$ |

This table basically tells us the pmf of $\overline{X}$: we just look at all the ways we can acheive each possible value of $\overline{X}$ (1, 1.5, 2, 2.5, 3, 3.5, and 4) and then add the probabilities together.

The pmf of $\overline{X}$ (which is often called the *sampling distribution* in

statistics) is thus

$$p_{\overline{X}}(x) = \begin{cases} 0.0025 & \text{if } x = 1 \\ 0.03 & \text{if } x = 1.5 \\ 0.15 & \text{if } x = 2 \\ 0.365 & \text{if } x = 2.5 \\ 0.39 & \text{if } x = 3 \\ 0.06 & \text{if } x = 3.5 \\ 0.0025 & \text{if } x = 4 \end{cases}$$

The pmf of $\overline{X}$ in the last example tells, for example, that if we were to pick two students from this university at random and average their GPA's, the probability the average would be 2 would be 15%. The example above is cheating a little bit because the whole point of the statistical theory we are about to develop will be to determine information about a population from a sample; in the example above we were given information about the entire population (the distribution of GPA's) and used this to find the probability a sample had a given value. To describe how to go the other way, we need to develop some more tools.

## 13.3 A note on notation

We will sometimes use upper- and lowercase letters to mean two slightly different things in the material to come. When discussing sample values in the abstract, we will use $X_i$ as a placeholder for the values to obtained in the sample, and use $x_i$ to mean a particular value we observed. That is, $X_i$ is a random variable, which we think of as a random value which will be obtained from our sample, and $x_i$ is a particular value. We extend this convention to statistics of our sample data. For example, $\overline{X}$ is a random variable which is a function of the random variables $X_1$, $X_2$, ..., $X_n$, but $\overline{x}$ is the average of the observed values $x_1$, $x_2$, ..., $x_n$.

# 14 Point Estimators

*I think that it is a relatively good approximation to truth – which is much too complicated to allow anything but approximations – that mathematical ideas originate in empirics.*

John von Neumann
*The Mathematician*

Much of statistics is about determining some parameter of the entire population, such as the average value or the variance of some quantity of interest, from samples. In this chapter we develop one technique for constructing these "point estimators," (estimates to population parameters) using *maximum likelihood estimators*. The basic idea is that we should be able to construct a statistic which for a given sample data gives us a "good" estimate of the population parameter we are interested in. In particular, we will treat the population parameter as an unknown variable, and consider following type of maximization problem: what choice of parameter maximizes the probability a random sample gives us the data we observed?

## 14.1 Point estimators in general

Suppose the population parameter we wish to estimate is denoted $\theta$. This is often the mean $\mu$ or the standard deviation $\sigma$, but it doesn't have to be. For example, with a binomial random variable where $n$ is known, we may be interested in the probability of success, $p$, and in that case our $\theta$ would be $p$. That is, $\theta$ is some parameter of the distribution of our random variables $X_i$ which we do not know and are trying to estimate. We will adopt the convention that for an unknown parameter $\theta$, the same parameter with a hat on it, $\widehat{\theta}$, is our estimate for the parameter. Notice this is a function of our sample data $X_1$, $X_2$, ..., $X_n$, and our goal is to find what this function must be.

One simple example of this is the sample mean. Given a collection of sample data $X_1$, ..., $X_n$, we might estimate the true value of the population mean (i.e., the value of $\mu = \mathbb{E}[X_i]$) using the sample mean,

$$\widehat{\mu} = \frac{X_1 + \cdots + X_n}{n}.$$

A very reasonable question to ask is whether this is really the best way to estimate $\mu$, and in this chapter we will prove that this really is the best we can do. (Here "best" means maximizing the probability of getting sample data $X_1$, ..., $X_n$. In other contexts or applications you may have a different idea of what "best" should mean.)

There are two common ways to construct these point estimators: by using *maximum likelihood estimators* and the *method of moments*. In class we will only discuss maximum likelihood estimators simply for the sake of time.

## 14.2 Maximum likelihood estimators

Suppose our population has some unknown parameter $\theta$ and we have a random sample (i.e., a collection of IID random variables) $X_1$, ..., $X_n$ from this population. The value $\theta$ influences the pdf (or pmf) of these random variables – that is, the function for the pmf or pdf depends on $\theta$ (for example, the parameter $\lambda$ in an exponential distribution). For this reason we will append a $\theta$ to the pmf/pdf of our random variables. For example, if the variables are discrete and $p(x)$ is the pmf, we will write the pmf as $p(x;\theta)$ to indicate that this pmf is also a function of $\theta$. Similarly, for continuous random variables we will write $f(x;\theta)$ for the pdf.

Since we have several random variables we are working with, $X_1$ through $X_n$, we should really be discussing the joint pmf (or pdf), so we will likewise append a $\theta$ to that function:

$$p(x_1, x_2, ..., x_n; \theta) = \Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

Notice, however, that since these random variables are assumed to be independent, we can write this joint pmf as the product of the individual pmf's:

$$p(x_1, x_2, ..., x_n; \theta) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta).$$

Our goal is to find the $\theta$ that maximizes this probability: we want to find a $\widehat{\theta}$ such that for any other choice of $\tilde{\theta}$,

$$p(x_1, x_2, ..., x_n; \widehat{\theta}) \geq p(x_1, x_2, ..., x_n; \tilde{\theta}).$$

Such a $\widehat{\theta}$, that maximizes the likelihood of a random sample being the given values $x_1$, $x_2$, ..., $x_n$ is called a ***maximum likelihood estimator***.

**Example 14.1.**
What percentage of American adults have the virus which may cause shingles? We of course can not take a blood sample of every adult American and test it to see if they have the *herpes zoster* virus, so instead we may take a sample of, say, 200.

Now imagine that we assign to each person in the population (to each American adult) a number: 0 if they do not have the virus, and 1 if they do. When we pick a random person from the population and test to see if they have the virus, we're really picking a random number that's either 0 or 1. For a given individual there's nothing random about this number (they either have the virus or they don't), but if we pick a random person we should interpret the number we get as random. That is, each member of our random sample gives us a Bernoulli random variable. The probability of "success" for such a random variable is exactly the proportion of people which have the virus, and this is our unknown parameter $\theta$. That is, the pmf of our random variables is

$$p(x;\theta) = \begin{cases} 1-\theta & \text{if } x = 0 \\ \theta & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

For our purposes it will be convenient to rewrite this pmf slightly as follows:

$$p(x;\theta) = \begin{cases} \theta^x \cdot (1-\theta)^{1-x} & \text{if } x = 0 \text{ or } 1 \\ 0 & \text{otherwise} \end{cases}$$

We're rewriting the pmf this way simply because we will want to consider the joint pmf for our 200 samples and this is a product of each of these pmf's. It will be easier to write down that product if we write the pmf as above.

The joint pmf of our 200 samples is thus

$$\begin{aligned}
&p(x_1, x_2, ..., x_n; \theta) \\
=&p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta) \\
=&\theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2} \cdots \theta^{x_n}(1-\theta)^{1-x_n} \\
=&\theta^{x_1+x_2+\cdots+x_n} \cdot (1-\theta)^{1-x_1+1-x_2+\cdots+1-x_n} \\
=&\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}
\end{aligned}$$

Now, given values of $x_1$, $x_2$, ..., $x_n$, we want to find the value of $\theta$ that maximizes the probability above. That is, we have a calculus problem: find the $\theta$ that maximizes $\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$.

To maximize this function we of course need to find the critical points, and so we have to solve the equation

$$\frac{d}{d\theta}\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = 0.$$

To make our lives a little bit easier, let's notice that since log is an increasing function, optimizing $\log(f(x))$ is just as good as optimizing $f(x)$. This is convenient because if we take a log, then the product above turns into a sum and is easier to differentiate. That is, we will actually maximize

$$\log\left(\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\right) = \log\left(\theta^{\sum_{i=1}^n x_i}\right) + \left(n - \sum_{i=1}^n x_i\right)\log(1-\theta).$$

Differentiating with respect to $\theta$ and setting this equal to zero we have

$$\left(\sum_{i=1}^n x_i\right) \cdot \frac{1}{\theta} + \left(n - \sum_{i=1}^n x_i\right) \cdot \frac{-1}{1-\theta} = 0.$$

Multiplying through by $\theta \cdot (1-\theta)$ to clear out the denominaotrs of the

fractions gives us

$$\begin{aligned}
&(1-\theta)\cdot\sum_{i=1}^{n}x_i-\theta\left(n-\sum_{i=1}^{n}x_i\right)=0\\
\implies&\sum_{i=1}^{n}x_i-\theta\sum_{i=1}^{n}x_i-n\theta+\theta\sum_{i=1}^{n}x_i=0\\
\implies&\sum_{i=1}^{n}x_i-n\theta=0\\
\implies&n\theta=\sum_{i=1}^{n}x_i\\
\implies&\theta=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{x_1+x_2+\cdots+x_n}{n}.
\end{aligned}$$

At this point we have a candidate for our maximum – we have a critical point. We still need to verify if this is in fact a maximum or not. To do that, let's consider the second derivative of the function we're trying to maximize:

$$\begin{aligned}
&\frac{d^2}{d\theta^2}\left[\log(\theta)\sum_{i=1}^{n}x_i+\log(1-\theta)\cdot\left(n-\sum_{i=1}^{n}x_i\right)\right]\\
=&\frac{d}{d\theta}\left[\frac{\sum_{i=1}^{n}x_i}{\theta}-\frac{n-\sum_{i=1}^{n}x_i}{1-\theta}\right]\\
=&\frac{d}{d\theta}\left[\theta^{-1}\sum_{i=1}^{n}x_i-(1-\theta)^{-1}\left(n-\sum_{i=1}^{n}x_i\right)\right]\\
=&-\theta^{-2}\sum_{i=1}^{n}x_i-(1-\theta)^{-2}\left(n-\sum_{i=1}^{n}x_i\right)\\
=&-\left[\frac{\sum_{i=1}^{n}x_i}{\theta^2}+\frac{n-\sum_{i=1}^{n}x_i}{(1-\theta)^2}\right].
\end{aligned}$$

Notice that as each $x_i$ is either zero or one, $n\geq\sum_{i=1}^{n}x_i$, and so the expression above is always negative. That is, our function is concave down everywhere, so our one critical point is in fact a global maximum.

What does all of this mean in context? Suppose we take a sample of 200 people which we test for the *herpes zoster* virus. The sum $\sum_{i=1}^{n}x_i$ simply counting the number of people which test positive for

the virus. Suppose this is 190 of our 200 samples. Then we should estimate that the parameter $\theta$ above is

$$\widehat{\theta} = \frac{190}{200} = 0.95$$

I.e., 95% of the population has the virus. Of all possible values of $\theta$, this is the one that maximizes the probability a sample of 200 individuals would have 190 which test positive.

The end result of the previous example is exactly what you would guess: we're just saying that if 95% of our sample has the virus, the most likely scenario is that 95% of the population has the virus. Notice that it is possible that some other proportion of the population has the virus and we just happened to pick a "bad" sample, but the work above shows the most likely scenario is that 95% of the population has the virus. For example, it is conceivable that only 1% of the population has the virus and somehow we just happened to pick a sample where 95% of the people in the sample had the virus. However, assuming the members of our sample are IID, this is very unlikely.

The function of $\theta$ above which we're trying to maximize is sometimes called the ***likelihood function*** since it tells us the probability (likelihood) of seeing the given data for a particular choice of $\theta$. In general, the likelihood function is a product of of the pmf (or pdf) for the individual random variables:

$$L(\theta) = p(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

for discrete random variables, and

$$L(\theta) = f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

for continuous random variables.

We want to maximize this function and so we must take a derivative. In the example above the function was nice enough that differentiating it directly could be done easily, but usually this is going to be a complicated derivative: it requires an $n$-fold product rule. To make this derivative easier, we take the logarithm to convert the product into a sum. This gives the

***log likelihood function***,

$$\log(L(\theta)) = \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) = \sum_{i=1}^{n} \log(f(x_i;\theta)).$$

Sometimes the distribution we are interested in involves multiple parameters. In such a situation our likelihood (and log likelihood) function becomes a function of several variables, and we must use the techniques of multivariable calculus to perform the optimization.

**Example 14.2.**
Suppose weights of adults is believed to be normally distributed. How can we estimate the mean $\mu$ and variance $\sigma^2$ from a sample of $n$ weights?

Our sample data in this case is given by $n$ IID random variables $X_1$, $X_2$, ..., $X_n$ where we are assuming each $X_i \sim N(\mu, \sigma)$, where $\mu$ and $\sigma$ are unknown. The likelihood function is thus

$$\begin{aligned} L(\mu,\theta) &= \prod_{i=1}^{n} f(x_i;\mu,\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{\sum_{i=1}^{n} \frac{-(x-\mu)^2}{2\sigma^2}}. \end{aligned}$$

The log likelihood function can now be written as

$$\begin{aligned} \log(L(\mu,\theta)) &= \log\left(\left(\sigma\sqrt{2\pi}\right)^{-n} e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= -n\log(\sigma\sqrt{2\pi}) + \log\left(e^{\sum_{i=1}^{n} \frac{-(x_i-\mu)^2}{2\sigma^2}}\right) \\ &= -n\log(\sigma) - n\log(\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned}$$

We want to maximize this function, so we need to find the critical points. Since this is a function of two variables, this requires us to

take the partial derivatives with respect to $\mu$ and $\sigma$:

$$\begin{aligned}\frac{\partial}{\partial\mu}\log(L(\mu,\sigma)) &= \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial\mu}(x_i-\mu)^2\\ &= \frac{-1}{2\sigma^2}\sum_{i=1}^{n}2(x_i-\mu)\cdot(-1)\\ &= \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}\end{aligned}$$

$$\frac{\partial}{\partial\sigma}\log(L(\mu,\sigma)) = \frac{-n}{\sigma}+\sigma^{-3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Now we solve the system of equations

$$\begin{aligned}\frac{\partial}{\partial\mu}\log(L(\mu,\sigma)) &= 0\\ \frac{\partial}{\partial\sigma}\log(L(\mu,\sigma)) &= 0\end{aligned}$$

The first equation we can directly solve for $\mu$:

$$\begin{aligned}
&\frac{\partial}{\partial \mu} \log(L(\mu, \sigma)) = 0 \\
\implies &\frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} = 0 \\
\implies &\sum_{i=1}^{n}(x_i - \mu) = 0 \\
\implies &\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0 \\
\implies &\sum_{i=1}^{n} x_i - n\mu = 0 \\
\implies &n\mu = \sum_{i=1}^{n} x_i \\
\implies &\mu = \frac{\sum_{i=1}^{n} x_i}{n}.
\end{aligned}$$

Solving the second equation for $\sigma^2$ gives

$$\begin{aligned}
&\frac{\partial}{\partial \sigma} \log(L(\mu, \sigma)) = 0 \\
\implies &\frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = 0 \\
\implies &\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma^3} = \frac{n}{\sigma} \\
\implies &\sum_{i=1}^{n}(x_i - \mu)^2 = n\sigma^2 \\
\implies &\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.
\end{aligned}$$

Keep in mind we already determined $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ above, however, and so we may write

$$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \frac{\sum_{i=1}^{n} x_i}{n}\right)^2}{n}.$$

To simplify notation, we may write $\overline{x}$ for $\frac{1}{n}\sum_{i=1}^{n} x_i$ and then the above becomes

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

Of course, $\sigma$ is simply the square root of this.

The above is telling us that if we measure weights of adults and record these weights as $x_1$, $x_2$, ..., $x_n$, assume that weights are normally distributed, then the most likely scenario is that the mean of the weights is

$$\widehat{\mu} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

and the variance is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

## 14.3 Biased and unbiased estimators

A point estimator $\widehat{\sigma}$ for a population parameter $\sigma$ is called ***unbiased*** if $\mathbb{E}[\widehat{\sigma}] = \sigma$, and is called ***biased*** otherwise. Keep in mind $\widehat{\sigma}$ is really a function of the observed data; we think of this data as being given by random variables, however, and so $\widehat{\sigma}$ is also a random variabe. Thus it makes sense to talk about the expected value of this random variable. Saying a point estimator is unbiased, then, means that if we looked at all possible inputs to the function (all possible sample data that could be observed) and averaged these together, that average would be the true value.

In our example where $X_i \sim \text{Bernoulli}(\theta)$, for example, we computed the maximum likelihood estimator of $\theta$ to be

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It is easy to see that this is an unbiased estimator by recalling that if $X_i \sim \text{Bernoulli}(\theta)$ then $\mathbb{E}[X_i] = \theta$ and applying basic properties of expected

values that we computed earlier:

$$\begin{aligned}
\mathbb{E}[\widehat{\theta}] &= \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\
&= \mathbb{E}\left[\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right] \\
&= \frac{1}{n}\mathbb{E}[X_1] + \frac{1}{n}\mathbb{E}[X_2] + \cdots + \frac{1}{n}\mathbb{E}[X_n] \\
&= \frac{1}{n}\theta + \frac{1}{n}\theta + \cdots + \frac{1}{n}\theta \\
&= n \cdot \frac{1}{n}\theta \\
&= \theta.
\end{aligned}$$

When $X_i \sim N(\mu, \sigma)$, we found in Example 14.2 that

$$\begin{aligned}
\widehat{\mu} &= \frac{\sum_{i=1}^n X_i}{n} \\
\widehat{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \widehat{\mu})^2}{n}.
\end{aligned}$$

Since $X_i \sim N(\mu, \sigma)$ means $\mathbb{E}[X_i] = \mu$, we easily see that $\widehat{\mu}$ is unbiased:

$$\begin{aligned}
\mathbb{E}[\widehat{\mu}] &= \mathbb{E}\left[\frac{\sum_{i=1}^n X_i}{n}\right] \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i]}{n} \\
&= \frac{\sum_{i=1}^n \mu}{n} \\
&= \frac{n\mu}{n} \\
&= \mu.
\end{aligned}$$

The manipulations for computing $\mathbb{E}[\widehat{\sigma^2}]$ are a little bit more involved. To explain how these manipulations work, let's first observe that for any random variable $X$, the variance is equal to $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, which we may rewrite as $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$ or $\mathbb{E}[X^2] = \sigma^2 + \mu^2$.

Now let's also note that by Corollary 12.10 we may write

$$\begin{aligned}\mathrm{Var}(\widehat{\mu}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}(X_i) \\ &= \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2 \\ &= n \cdot \frac{1}{n^2}\sigma^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Combining these observations together we have

$$\mathbb{E}[\widehat{\mu}^2] = \mathrm{Var}(\widehat{\mu}) + \mathbb{E}[\widehat{\mu}]^2 = \frac{\sigma^2}{n} + \mu^2.$$

This last observation is the key trick for computing $\mathbb{E}[\widehat{\sigma}^2]$. We begin by writing out the definition and doing some basic algebra:

$$\begin{aligned}\mathbb{E}[\sigma^2] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i - \widehat{\mu})^2}{n}\right] \\ &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i^2 - 2X_i\widehat{\mu} + \widehat{\mu}^2)}{n}\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2X_i\widehat{\mu} + \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\widehat{\mu}\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\cdot n\widehat{\mu}^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\widehat{\mu}\cdot\widehat{\mu} + \widehat{\mu}^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \widehat{\mu}^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\widehat{\mu}^2\right].\end{aligned}$$

Now we apply the observations above:

$$\begin{aligned}
\mathbb{E}[\sigma^2] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\widehat{\mu}^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2+\mu^2\right) - \left(\frac{\sigma^2}{n}+\mu^2\right) \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n\sigma^2 - \sigma^2}{n} \\
&= \frac{\sigma^2(n-1)}{n} \\
&= \sigma^2\frac{n-1}{n}.
\end{aligned}$$

Notice that $\frac{n-1}{n} \neq 1$, and so $\mathbb{E}[\widehat{\sigma^2}] \neq \sigma^2$, and so this is a biased estimator.

Having a biased estimator is not necessarily "bad" (the name *biased* might have some negative connotations, but mathematically there is nothing wrong with a biased estimator), although it is soemtimes desirable to have unbiased estimators. Though our maximum likelihood estimator $\widehat{\sigma^2}$ is biased, we can actually perform one simple manipulation to get an unbiased estimator.

Given $n$ IID random variables $X_1$, $X_2$, ..., $X_n$ where $\widehat{\mu}$ is the maximum likelihood estimator for $\mathbb{E}[X]$, we define the ***sample variance***, denoted $S^2$, as

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \widehat{\mu}\right)^2}{n-1}.$$

That is, $S^2$ is almost the same as our $\widehat{\sigma^2}$ from earlier, except we divide the sum by $n-1$ instead of $n$. The claim is that this minor modification will make $S^2$ an unbiased estimator.

**Proposition 14.1.**
*$S^2$ is an unbiased point estimator for the variance* $\mathrm{Var}(X_i)$ *for a collection of $n$ IID random variables, $X_1$ through $X_n$.*

*Proof.*
We must check thtat $\mathbb{E}[S^2] = \sigma^2$. First we write out the definition of $S^2$ in expected value, then factor out the $n-1$ and move it to the other side, and expand the square in the sum.

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i - \widehat{\mu})^2}{n-1}\right]$$
$$\implies (n-1)\mathbb{E}[S^2] = \mathbb{E}\left[\sum_{i=1}^{n}\left(X_i^2 - 2X_i\widehat{\mu} + \widehat{\mu}^2\right)\right].$$

Now we use our properties of expected value to split up the sum and factor out any constants,

$$\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i^2 - 2X_i\widehat{\mu} + \widehat{\mu}^2\right)\right]$$
$$= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[\widehat{\mu}]\mathbb{E}\left[\sum_{i=1}^{n}X_i\right] + \mathbb{E}\left[\sum_{i=1}^{n}\widehat{\mu}^2\right].$$

Now we rewrite each $\mathbb{E}[X_i^2]$ as $\sigma^2 + \mu^2$, which we had observed earlier. We also note that since $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}X_i$, the $\mathbb{E}\left[\sum_{i=1}^{n}X_i\right]$ factor in the middle term can be rewritten as $n\mathbb{E}[\widehat{\mu}]$. Using these observations we write

$$\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[\widehat{\mu}]\mathbb{E}\left[\sum_{i=1}^{n}X_i\right] + \mathbb{E}\left[\sum_{i=1}^{n}\widehat{\mu}^2\right]$$
$$= \sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - 2n\mathbb{E}[\widehat{\mu}^2] + n\mathbb{E}[\widehat{\mu}^2].$$

Combining like-terms this becomes

$$\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - 2n\mathbb{E}[\widehat{\mu}^2] + n\mathbb{E}[\widehat{\mu}^2] = n(\sigma^2 + \mu^2) - n\mathbb{E}[\widehat{\mu}^2].$$

Now recalling $\mathrm{Var}(\widehat{\mu}) = \sigma^2/n$ and that for any random variable $X$ we may write $\mathbb{E}[X^2] = \mathrm{Var}(X) + \mathbb{E}[X]^2$, the $\mathbb{E}[\widehat{\mu}^2]$ above can be rewritten

to obtain

$$\begin{aligned} n(\sigma^2 + \mu^2) - n\mathbb{E}[\widehat{\mu}^2] &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\ &= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 \\ &= n\sigma^2 - \sigma^2 \\ &= \sigma^2(n-1) \end{aligned}$$

At this point we have shown $(n-1)\mathbb{E}[S^2] = \sigma^2(n-1)$. Finally, dividing both sides by $n-1$ proves the proposition. □

## 14.4 Practice problems

**Problem 14.1.**
Suppose $X_1$, $X_2$, ..., $X_n$ are IID random variables where each $X_i$ is a uniform random variable $X_i \sim \text{Uni}[-\theta, \theta]$, and $\theta$ is unknown. I.e., the pdf of each $X_i$ is

$$f(x) = \begin{cases} \frac{1}{2\theta} & \text{if } -\theta \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Compute the maximum likelihood estimator $\widehat{\theta}$ of $\theta$.

**Problem 14.2.**
Suppose the lifetime of a certain type of light bulb is exponentially distributed with unknown parameter $\lambda$. Suppose a random sample of $n$ light bulbs is taken, and the lifetime of the $i$-th lightbulb is recorded as $x_i$ (in hours, say). Find a formula for the maximum likelihood estimator $\widehat{\lambda}$ of $\lambda$.

**Problem 14.3.**
The *Gamma function* is defined for $x > 0$ to be

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt.$$

One key property of the $\Gamma$ function is that for every $x > 0$, $\Gamma(x+1) = x \cdot \Gamma(x)$.

The *Gamma distribution* is a continuous distribution which depends on two parameters, $\alpha > 0$ and $\beta > 0$, and has the following pdf:

$$f(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\Gamma$ is Gamma function defined above. If a random variable $X$ has the above pdf, we will write $X \sim \text{Gamma}(\alpha, \beta)$.

(a) Suppose $X \sim \text{Gamma}(\alpha, \beta)$. Compute the expected value $\mathbb{E}[X]$. Your answer should be a function of $\alpha$ and $\beta$ and *should not* have any $\Gamma$'s!

(Hint: Write out the definition of the expected value for $X$ using the pdf of the Gamma distribution, pull out any terms that don't depend on $x$, and then use the definition of the $\Gamma$ function, and the property of the $\Gamma$ function mentioned above.)

(b) Compute $\mathbb{E}[X^2]$. Again, your answer should be a function of $\alpha$ and $\beta$, but should not have any $\Gamma$'s.

(Hint: Write out the definition of $\mathbb{E}[X^2]$. Then multiply and divide by the same number in such a way that, after pulling some constants out of the integral, you have the formula for the integral of the pdf of a random variable with distribution $\Gamma(\alpha + 2, \beta)$. Use the fact that pdf's integrate to 1, and then use a property of the $\Gamma$ function defined above.)

(c) Suppose $X_1$, $X_2$, ..., $X_n$ are IID random variables where each $X_i \sim \text{Gamma}(\alpha, \beta)$, but where $\alpha$ and $\beta$ are unknown. Use the method of moments to construct estimators for $\alpha$ and $\beta$.

# 15

# *Confidence Intervals*

*"Reeling and Writhing, of course, to begin with," the Mock Turtle replied, "and then the different branches of Arithmetic: Ambition, Distraction, Uglification, and Derision."*

LEWIS CARROLL
*Alice's Adventures in Wonderland*

## 15.1 Idea of a confidence interval

The point estimators we've discussed are, of course, only estimates of the values we care about, and so a reasonable question to ask is how good of an estimate are they? A related question is can we construct estimates that have some pre-determined degree of accuracy?

We can do this by constructing "confidence intervals." The idea being that we collect sample data and from this data compute a range of possible values for the parameter of interest.

For example, suppose we want to know the average height of students at IU. We may reasonably suppose heights are normally distributed with standard deviation, say, $\sigma = 3$ inches. (More on assuming we know $\sigma$ later.) We may not know the true value of the mean $\mu$, but notice that whatever $\mu$ happens to be, we can transform the sample mean $\overline{X}$ for some sample of $n$ heights to the standard normal. The key to doing this is the following theorem:

**Theorem 15.1.**
*If $X_1$, $X_2$, ..., $X_n$ are IID random variables which are each $N(\mu, \sigma)$, then their sample mean $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is a normal random variable: $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$.*

**Remark.**
Notice that the theorem above is very similar to our alternative version

of the central limit theorem, Theorem 12.13. The difference between this theorem and the central limit theorem is that Theorem 15.1 says $\overline{X}$ is *exactly* a normal random variable, and Theorem 12.13 says $\overline{X}$ is approximately normal. However, Theorem 12.13 applies for *any* distribution, whereas Theorem 15.1 requires the $X_i$ are all normal.

Since $\overline{X} \sim N(\mu, \sigma/\sqrt{n})$,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is the standard normal. Notice that for the standard normal we can determine the interval where 95% of the outputs live (we can't really compute this by hand, but can look up approximations done on a computer):

$$\Pr(-1.96 < Z < 1.96) = 0.95.$$

What this means for our $X_1$, $X_2$, ..., $X_n$ random variables is that 95% of the itme we compute a sample mean $\overline{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ where each $x_i$ is the value of the random variable $X_i \sim N(\mu, \sigma)$, the corresponding $z$-value,

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}},$$

we'd obtain a value in the interval $(-1.96, 1.96)$. We can manipulate this to get an interval for $\mu$. In particular, let's suppose there are $n = 100$ samples, and recall we assumed $\sigma = 3$. Then 95% of the time, the value

$$\frac{\overline{x} - \mu}{3/10}$$

is between $-1.96$ and 1.96. That is,

$$\begin{aligned}
&-1.96 < \frac{\overline{x} - \mu}{3/10} < 1.96 \\
\implies &\frac{10}{3} \cdot (-1.96) < \overline{x} - \mu < \frac{10}{3} \cdot 1.96 \\
\implies &-\overline{x} - 1.96 \cdot \frac{10}{3} < -\mu < -\overline{x} + 1.96 \cdot \frac{10}{3} \\
\implies &\overline{x} - 1.96 \cdot \frac{10}{3} < \mu < \overline{x} + 1.96 \cdot \frac{10}{3} \\
\implies &\overline{x} - 6.53 < \mu < \overline{x} + 6.3
\end{aligned}$$

That is, for a given collection of sample data $x_1$, $x_2$, ..., $x_n$, 95% of the time the true value of $\mu$ will be in the range $(\overline{x} - 6.53, \overline{x} + 6.53)$. For example. if

our sample data is such that the sample mean is $\overline{x} = 68$, then we are 95% confident the true mean $\mu$ is in the interval $(61.47, 74.53)$.

Notice that the true mean $\mu$ is not a random quantity; it is some number, whatever it happens to be, that does not change. We're trying to estimate that value from sample data, however, and this sample data may change from one sample to the next. That is, the mean doesn't change, but the interval we compute does. For all possible intervals we could construct in the manner above, 95% of them will contain the true mean.

## 15.2 Confidence intervals in general

In general, if $X_1$, ..., $X_n$ are IID random variables which are normal with unkown mean $\mu$ and known standard deviation $\sigma$, then from $n$ sample values $x_1$, $x_2$, ..., $x_n$, the 95% confidence interval for $\mu$ is

$$\left(\overline{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right).$$

We could similarly compute 99% or 90% confidence intervals, or 87.2% confidence intervals. The key for doing this is to know what the corresponding interval for the standard normal would be. To make this precise, recall that for each $\alpha \in (0, 1)$ we defined the ***z-critical value*** $z_\alpha$ to be the value of $z_\alpha$ such that $\Pr(Z \geq z_\alpha) = \alpha$. This tells us that $100 \cdot \alpha\%$ of the output of the standard normal is to the left of $z_\alpha$. For our purposes we want to find the middle 95%, or 99%, or 90%, or 87.2%. Because the standard normal's pdf is symmetric about zero, we thus take the percentage we care about, cut it in half, and use that as our $\alpha$.

For example, to find the interval $(-\zeta, \zeta)$ containing 90% of the data for the standard normal, we need that 10% of the data lives outside the interval. By symmetry, 5% will be to the right of our interval and 5% will be to the left. This means we need to take $\zeta$ to be the value $z_{0.05}$ which we can look up is 1.645:

$$\Pr(-1.645 < Z < 1.645) = 0.9.$$

Standardizing a sample mean $\overline{X}$ to get the standard normal $Z$, we then

work backwards to obtain the 90% confidence interval:

$$\begin{aligned} & -1.645 < Z < 1.645 \\ \implies & -1.645 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.645 \\ \implies & -1.645\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.645\frac{\sigma}{\sqrt{n}} \\ \implies & -\overline{X} - 1.645\frac{\sigma}{\sqrt{n}} < -\mu < -\overline{X} + 1.645\frac{\sigma}{\sqrt{n}} \\ \implies & \overline{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.645\frac{\sigma}{\sqrt{n}} \end{aligned}$$

and so the 90% confidence interval of $\mu$ is

$$\left(\overline{X} - 1.645\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.645\frac{\sigma}{\sqrt{n}}\right).$$

## 15.3 The effect of sample size

In general, the $100(1-\alpha)\%$ confidence interval for $\mu$ is

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

Notice that for a fixed sample size, asking for more confidence – 90%, 95%, 99%, etc. – makes the interval larger since more confidence means larger $z_{\alpha/2}$-values.

What if we want to instead make the interval smaller? From the formula

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

we see that to make the interval smaller, we need to increase the number of samples, $n$. Note that the width of the $100(1-\alpha)\%$ confidence interval is

$$2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

If we want this width to be no bigger than some given value $w$, how large should $n$ be? This is a simple algebra problem:

$$\begin{aligned} & 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < w \\ \implies & 2z_{\alpha/2}\frac{\sigma}{w} < \sqrt{n} \\ \implies & n \geq \left(2z_{\alpha/2}\frac{\sigma}{w}\right)^2. \end{aligned}$$

So if $\sigma = 3$, for instance, and we want a 99% confidence interval of width at most $1/2$, then the number of required samples $n$ is

$$\begin{aligned} n > \left(2\frac{z_{0.005} \cdot 3}{1/2}\right)^2 &= (12 z_{0.005})^2 \\ &= (12 \cdot 2.58)^2 \\ &= 958.5 \end{aligned}$$

That is, we require at least 959 samples to have a confidence interval of width at most $1/2$. We'll see an example of this in a moment, but first we should address the elephant in the room.

## 15.4 What if the distribution is not normal?

Previously we had assumed that the underlying distribution of our data was normal, but this need not be the case. For example, suppose each member of our population is classified as having some attribute or not, which we denote by assigning the member a 1 or 0. Then each member of our sample gives us a Bernoulli random variable $X_i \sim \text{Bernoulli}(p)$, where $p$ represents the proportion of the population having that attribute. In the case of a Bernoulli random variable with parameter $p$, we know the variance is given by $p(1-p)$. The central limit theorem then tells us that for a sufficiently large sample size $n$, the sample mean $\overline{X}$ is a normal random variable with mean $\mu = p$ and standard deviation $\sqrt{p(1-p)}/n$. If $p$ is unknown, we can build a confidence interval for $p$ use the same techniques as before.

Since $\overline{X}$ is approximately $N(p, \frac{p(1-p)}{n})$, the random variable

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately the standard normal. Thus for any $\alpha \in (0, 1)$,

$$\Pr\left(-z_{\alpha/2} < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

Now we want to manipulate the inequalities

$$-z_{\alpha/2} < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}$$

to get an interval centered around $p$. This takes a little bit of algebraic work, so we'll state the final result as a proposition and relegate all of that algebra to the proof of the proposition.

**Proposition 15.2.**
*The* $100(1-\alpha)\%$ *confidence interval for the proportion* $p$ *of the population having some given attribute is estimated, using a sample of large enough size* $n$ *with where the proportion of members of the same having the attribute is* $\overline{x}$*, to be*

$$\left( \frac{\overline{x} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} - z_{\alpha/2} \sqrt{\frac{\frac{\overline{x}(1-\overline{x})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}{1 + \frac{{z_{\alpha/2}}^2}{n}}}, \frac{\overline{x} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} + z_{\alpha/2} \sqrt{\frac{\frac{\overline{x}(1-\overline{x})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}{1 + \frac{{z_{\alpha/2}}^2}{n}}} \right)$$

*Proof.*
We continue where we left off above with the inequality

$$-z_{\alpha/2} < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}.$$

Notice this may be written as

$$\left| \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| \leq z_{\alpha/2}.$$

We now square this to write

$$\frac{(\overline{X} - p)^2}{\left(\frac{p(1-p)}{n}\right)} \leq z_{\alpha/2}.$$

Now we do some algebra:

$$\begin{aligned}
&\frac{(\overline{X}-p)^2}{\left(\frac{p(1-p)}{n}\right)} \leq z_{\alpha/2} \\
\implies & n(\overline{X}-p)^2 \leq z_{\alpha/2}^2 p(1-p) \\
\implies & n\overline{X}^2 - 2n\overline{X}p + np^2 \leq z_{\alpha/2}^2 \cdot (p-p^2) \\
\implies & (n+z_{\alpha/2}^2)p^2 + (-z_{\alpha/2}^2 - 2n\overline{X})p + n\overline{X}^2 \leq 0.
\end{aligned}$$

Now we complete the square:

$$\begin{aligned}
& p^2 + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{n+z_{\alpha/2}^2}p + \frac{n\overline{X}}{n+z_{\alpha/2}^2} \leq 0 \\
\implies & p^2 + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{n+z_{\alpha/2}^2}p \leq \frac{-n\overline{X}}{n+z_{\alpha/2}^2} \\
\implies & p^2 + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{n+z_{\alpha/2}^2}p + \left(\frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2 \leq \frac{-n\overline{X}}{n+z_{\alpha/2}^2} + \left(\frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2 \\
\implies & \left(p + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2 \leq \frac{-n\overline{X}}{n+z_{\alpha/2}^2} + \left(\frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2
\end{aligned}$$

Finally, taking the positive and negative square roots we obtain the inequalities

$$\begin{aligned}
p + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2} &\leq \sqrt{\frac{-n\overline{X}}{n+z_{\alpha/2}^2} + \left(\frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2} \\
-\left(p + \frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right) &\geq -\sqrt{\frac{-n\overline{X}}{n+z_{\alpha/2}^2} + \left(\frac{-z_{\alpha/2}^2 - 2n\overline{X}}{2n+2z_{\alpha/2}^2}\right)^2}
\end{aligned}$$

Finally, simplifying the terms in the expressions above and solving the inequalities for $p$ gives us the result:

$$\frac{\overline{x} + \frac{z_{\alpha/2}^2}{2n}}{1+\frac{z_{\alpha/2}^2}{n}} - z_{\alpha/2}\sqrt{\frac{\frac{\overline{x}(1-\overline{x})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}{1+\frac{{z_{\alpha/2}}^2}{n}}} \leq p \leq \frac{\overline{x} + \frac{z_{\alpha/2}^2}{2n}}{1+\frac{z_{\alpha/2}^2}{n}} + z_{\alpha/2}\sqrt{\frac{\frac{\overline{x}(1-\overline{x})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}{1+\frac{{z_{\alpha/2}}^2}{n}}}$$

□

**Example 15.1.**
A researcher is interested in the proportion of people using illegal drugs. To estimate this proportion, the researcher has 500 people anonymously fill out a survey indicating whether they had used an illegal drug in the last twelve months or not. Supposing 47 people indicated they had used illegal drugs in the last year, construct a 95% confidence interval for the proportion of the population using an illegal drug in the last twelve months.

We are constructing a 95% confidence interval, so $\alpha = 0.05$, and $\alpha/2 = 0.025$. The corresponding $z$-critical value is $z_{0.025} = 1.96$. In our sample $n = 500$ and the proportion of people using an illegal drug is $\overline{x} = 47/500 = 0.094$, so $1 - \overline{x} = 0.906$. Plugging all of this into the formula for the confidence interval above, the endpoints of the our confidence interval are

$$\frac{\overline{x} + \frac{z_{\alpha/2}^2}{2n}}{1 + \frac{z_{\alpha/2}^2}{n}} \pm z_{\alpha/2}\sqrt{\frac{\frac{\overline{x}(1-\overline{x})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}{1 + \frac{{z_{\alpha/2}}^2}{n}}} = 0.0971 \pm 0.0258.$$

Our interval is $(0.0713, 0.1229)$. That is, we estimate with 95% confidence that between 7.13% and 12.29% of the population has used an illegal drug, based on a sample of size 500 where 9.4% of the people in the sample used an illegal drug.

## 15.5 What if the distribution is unknown?

In the previous example we assumed the distribution of our data was Bernoulli, which made sense as each member of the population was assigned a 0 or 1. In that case we were then able to use the fact that the variance $\sigma^2$ of a Bernoulli random variable with parameter $p$ was $p(1-p)$. In general, the numbers we assign to members of the population may not be simply zeros or ones, and so we may not be able to assume the random variables are Bernoulli. In fact, we may not know what the distribution of our

random variables is at all! I.e., we may only have sample data (a collection of numbers) and no knowledge of the underlying distribution. Ultimately, if we 're trying to construct a confidence interval for the mean $\mu$ of the population, the central limit theorem allows us to side-step the issue that the distribution is not known. However, we still need to know the standard deviation of the distribution. When we knew the distribution, as in the Bernoulli case, we may have a nice formula for this standard deviation, but what if we don't know the distribution and so have no such formula?

Even though the true standard deviation is unknown, we saw in the last chapter how to construct point estimators. In particular, we saw that the sample variance $S^2$ is an unbiased estimator for the variance. Taking the square root of this we have an estimate for the standard deviation: the ***sample standard deviation*** $S$ is

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

So we will compute the sample standard deviation $S$ and use this in place of the $\sigma$ in our confidence intervals above.

In general we may not know the distribution of our data: we may just know the sample values $x_i$ and have no knowledge of how the random variables $X_i$ are supposed to be distributed. As far as confidence intervals are concerned, however, we don't actually need to distribution of the individual $X_i$'s: we need the distribution of the sample mean $\overline{X}$. The central limit theorem, however, tells us that for *any* distribution, for a sufficiently large number of samples, sample mean $\overline{X}$ is "approximately" normal with mean $\mu$ and standard deviation $\sigma/n$. Replacing $\sigma$ with our approximation, the sample standard deviation $S$, we have that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is approximately the standard normal.

Putting all of this together, if we have sample values $x_1$, $x_2$, ..., $x_n$, we compute the sample mean $\overline{x}$ as

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

and the sample standard deviation $s$ as

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

and then estimate the 95% confidence interval of $\mu$ as

$$\left(\overline{x} - 1.96\frac{s}{\sqrt{n}}, \overline{x} + 1.96\frac{s}{\sqrt{n}}\right),$$

or the 90% confidence interval is estimated as

$$\left(\overline{x} - 1.645\frac{s}{\sqrt{n}}, \overline{x} + 1.645\frac{s}{\sqrt{n}}\right).$$

In general, the $100(1-\alpha)\%$ confidence interval for the population mean $\mu$ is approximately

$$\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

**Example 15.2.**
Repeating Example 15.1 but using the sample standard deviation gives us a confidence interval of $(0.0711, 0.1169)$ and so we instead estimate the number of individuals int he population having used an illegal drug is between 7.11% and 11.69%.

**Example 15.3.**
The BRCA1 and BRCA2 genes are thought to be related to tumor suppression, and mutations in these genes are believed to be related to a higher risk for developing breast cancer. An oncologist is interested in knowing what percentage of breast cancer patients have a mutation in these genes, and so performs genetic testing on 923 randomly selected breast cancer patients to determine if they have a BRCA mutation or not. Of the 923 patients, 74 have the mutation.

(a) Construct a 95% confidence interval for the proportion of all breast cancer patients which have the mutation.

(b) If the oncologist wanted a 99% confidence interval whose width was at most two percentage points (i.e., the interval is of the form $(\overline{X} - 0.01, \overline{X} + 0.01)$), how many patients would need to be tested?

---

(a) We construct a 95% confidence interval by simply plugging our given data into the formula. Note here that the sample mean is $\frac{74}{923} = 0.081$. For the standard deviation we are required to use the sample standard deviation which is

$$\begin{aligned} S &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n}(x_i - {}^{74}\!/_{923})^2}{922}} \\ &= \sqrt{\frac{74 \cdot (1 - {}^{74}\!/_{923})^2 + 849 \cdot (0 - {}^{74}\!/_{923})^2}{922}} \\ &\approx \sqrt{\frac{62.6099}{922}} \\ &\approx 0.2717 \end{aligned}$$

Using this in the same formula as the confidence interval for the previous problem, expect with $S$ in place of $\sigma$, we have a 95% confidence interval of $(0.0796, 0.0808)$, or, as percentages, $(7.96\%, 8.08\%)$.

(b) Recall that the $100(1-\alpha)\%$ confidence interval for the mean $\mu$ of a normal population with standard deviation $\sigma$ taken from a sample of size $n$ with sample mean $\overline{X}$ has width $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. When the standard deviation $\sigma$ is unknown, we approximate it with the sample standard deviation $S$. We want this interval of width at most 0.02 when $\alpha = 0.01$. Using $z_{0.005} = 2.58$ and the sample standard deviation $S = 0.2717$ from part (a) (which is only an estimate since we don't know the true standard deviation), we have the following:

$$\begin{aligned} & 2 \cdot 2.58 \cdot \frac{0.2717}{\sqrt{n}} < 0.02 \\ \implies & 2 \cdot 2.58 \cdot \frac{0.2717}{0.02} < \sqrt{n} \\ \implies & n > \left(2 \cdot 2.58 \cdot \frac{0.2717}{0.02}\right)^2 \approx 4913.81 \end{aligned}$$

Thus the oncologist would require a sample of at least 4914 patients.

## 15.6 Practice problems

**Problem 15.1.**
Suppose GPA's at a university are normally distributed with standard deviation $\sigma = 0.5$. You conduct an experiment where ten randomly selected students are asked what their GPA's are, and receive the following answers:

$$3.2,\ 3.4,\ 2.5,\ 3.7,\ 2.5,\ 3.1,\ 1.7,\ 3.9,\ 2.8,\ 1.2$$

Construct a 95% confidence interval for the true mean of GPA's at the university.

**Problem 15.2.**
Suppose heights of players in the NBA, measured in inches, are normallly distributed with unknown mean $\mu$ and standard deviation $\sigma = 1.5$. A sample of 16 random players yields a sample mean height of 78. Construct a 95% confidence interval for the population mean height $\mu$. (Use $z_{0.025} = 1.96$.)

# 16 Hypothesis Testing

*Hanging on in quiet desperation is the English way*
*The time is gone, the song is over,*
*Thought I'd something more to say.*

---

Pink Floyd
*Time*

## 16.1 Idea and motivating example

Hypothesis testing is about determining if there exists enough evidence (e.g., data from a sample) to replace a default hypothesis with an alternative one. The common analogy is the legal system where (ideally) someone accused of a crime is considered innocent by default unless there is enough evidence to make it likely they are guilty.

The "default" hypothesis is called the ***null hypothesis*** and is often denoted $H_0$. The alternative being tested is called the ***alternative hypothesis*** and is denoted $H_a$. Our goal is to determine if there's significant data taken from a sample to reject the null hypothesis or not.

For example, a doctor may be interested in the effectiveness of a new type of drug for high cholesterol. Suppose that among a population of patients with high cholesterol, the amount of choloesterol in a person's bloodstream, measured in milligrams per deciliter, is normally distributed with a mean of 260 and standard deviation of 30. The doctor may give a sample of ten patients the new drug for thirty days, record their cholesterol after being on the medication for thirty days, and then determine the sample mean of the cholesterol is 245 with a satndard deviation of 20. We want to determine if this is sufficient evidence to determine if the medication is effective.

At the start of this experiment we do not yet know if the medication is effective or not; *a priori* we have no reason to believe the medication is effective, so we will assume it is not effective unless there is sufficient evidence to say otherwise. That is, our null hypothesis is that the medication is not effective at controlling cholesterol levels, and the alternative hypothesis is that the medication is effective. To help us determine if the medication is effective or not, we are considereing the sample mean of cholesterol levels in patients that have been taking the medication for thirty days. This sample

mean is called a ***test statistic***, and it is what will ultimately tell us if there is sufficient evidence to reject the null hypothesis or not.

If the medicine is effective, we expect cholesterol of people in the sample to be lower than the cholesterol levels from people in our population of high cholesterol patients. But how low is low enough to conclude the medicine works? To answer this we may want to consider the possibility that by random chance the people in our sample simply had slightly lower-than-average cholesterol levels, among the people in our high cholesterol population. That is, we know on average the cholesterol levels are 260 mg/dL – but some people will be above this average and some will be below. Is it likely that just by random chance we happened to select a sample that is below average?

Answering this question in general might not be easy, but keep in mind we know the data is normally distributed in this example. Since we have a non-standard normal we can standardize it and this will give us some information about how likely it is to randomly select patients which will give us a sample mean below the population mean. Transforming $N(260, 30)$ to $Z \sim N(0, 1)$, the value of 245 becomes

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{245 - 260}{30/\sqrt{10}} = -1.58.$$

How likely is it our sample is just lower than average without the medication? Well, the above tells us a random sample of ten patients would have a sample mean of 245, just by random chance, with probability

$$\Phi(-1.58) \approx 0.057.$$

That is, there's only a 5.7% chance we would find such a sample, and so this may be good enough for the doctor to conclude the medicine is effective.

Normally in doing hypothesis testing we decide before-hand how strong the data needs to be for us say whether or not the null hypothesis should be rejected. For example, we might agree to reject the null hypothesis only if there is less than 10% chance our collected data would contradict the null hypothesis by random chance; or we may decide there should be only a 1% chance of such data. This "cutoff" value is called the ***significance level*** of the test and is often denoted $\alpha$. Once we know what $\alpha$ is, we can determine what the cutoff values are for the ***rejection region***, the range of $z$-values where we would reject the null hypothesis.

Notice whether we reject the null hypothesis or not depends on what significance level we want. In the case of the cholesterol medication above, we would reject the null hypothesis at the 10% significance level, but not at the 5% significance level.

## 16.2 Examples

**Example 16.1.**
A chocolatier claims an average box of their chocolates weigh 368g with a standard deviation of 10g. A sample of 49 boxes has sample mean $\overline{x} = 364g$. Test the hypothesis the mean weight less than the claimed $368g$ with a significance level of $\alpha = 0.05$.

Here the null hypothesis is

$$H_0 : \mu = 368$$

and the altneriatve is

$$H_a : \mu < 368.$$

For a significance level of $\alpha = 0.05$, our cutoff region is the given by the value of $\zeta$ such that $\Pr(Z < \zeta) = 0.05$. I.e., fifth percentile of the standard normal, which we can look up is $\zeta = -1.645$.

So, when we compute our test statistic in a moment, if $z < -1.645$, then we will reject the null hypothesis. If $z \geq -1.645$, then we will fail to reject the null hypothesis. What this means is that, after normalizing, there is only a 5% chance our computed $z$ value would be less than $-1.645$.

Now we compute our test statistic,

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{349 - 368}{10/sqrt49} = -2.8.$$

Thus we reject the null hypothesis: there is sufficient evidence to conclude the average weight of the chocolatier's boxes of chocolates is less than 368 grams.

**Example 16.2.**
A certain type of car is advertised as being able to average fifty miles per gallon, and the company claims the standard deviation is $\sigma = 5$. A sample of 20 cars are driven and their average mileage determined to be 48 mpg. Test the hypothesis the average mileage of these cars is

not 50 mpg with significance level $\alpha = 0.05$.

The null and alternative hypotheses are

$$H_0 : \mu = 50$$
$$H_a : \mu \neq 50.$$

For our cutoff values, since we're only testing $\mu \neq 50$ and not $\mu < 50$, we need to consider the case that $\mu < 50$ or $\mu > 50$. With a significance level of 5%, this means we want to have a 2.5% chance of $\mu < 50$ and a 2.5% chance of $\mu > 50$. That is, after standardizing, we want to find the value of $\zeta$ such that $\Pr(-\zeta < Z < \zeta) = 0.95$ so $\Pr(Z < -\zeta \text{ or } Z > \zeta) = 0.05$. By the symmetry of the pdf of the standard normal, this means we need to find the $\zeta$ such that $\Pr(Z < \zeta) = 0.025$, which we can look up means $\zeta = 1.96$. Thus we will reject the null hypothesis if $z < -1.96$ or $z > 1.96$. (These values correspong to a 2.5% chance that by "dumb luck" the real mean is 50 mpg but our samples were above or below the mean.)

We compute our test statistic,

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{48 - 50}{5/\sqrt{20}} = -1.79.$$

This value is *not* in our rejection region, so we fail to reject the null hypothesis. That is, there is not sufficient evidence to reject the car company's claim their cars average 50 mpg.

## 16.3 Tails, rejection regions, and $P$-values

Usually when performing the null hypothesis testing, the null hypothesis takes the form $H_0 : \mu = \mu_0$. I.e., the claim we are testing is that the population mean $\mu$ is some given value $\mu_0$. The alternative hypothesis then takes one of the following forms:

- $H_a : \mu > \mu_0$,
- $H_a : \mu < \mu_0$, or
- $H_a : \mu \neq \mu_0$.

These situations are called ***upper-tailed***, ***lower-tailed***, and ***two-tailed tests***, respectively.

The rejection region is slightly different in each case. For a fixed $\alpha$ we have the following rejection regions:

**Upper-tailed tests** For an upper-tailed test, we reject the null hypothesis if $z > z_\alpha$.

**Lower-tailed tests** For a lower-tailed test, we reject the null hypothesis if $z < -z_\alpha$.

**Two-tailed tests** For a two-tailed test, we reject the null hypothesis if either $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$.

There is an alternative way of determining whether we should reject the null hypothesis or not without computing the rejection region using a number called a $P$ value. The ***P-value*** is a number which indicates the probability of our test statistic being at least as contradictory to $H_0$ as the observed data – notice that "at least as contradictory" depends on whether our hypothesis test is upper-tailed, lower-tailed, or two-tailed. The $P$-value in each case is computed as follows:

$$P = \begin{cases} 1 - \Phi(z) & \text{for an upper-tailed test} \\ \Phi(z) & \text{for a lower-tailed test} \\ 2(1 - \Phi(|z|)) & \text{for a two-tailed test} \end{cases}$$

We will reject the null hypothesis if $P$ is less than our significance level $\alpha$.

In each of our three examples thus far (the cholesterol medication, the chocolatier, and the cars), our $P$ values are as follows:

- Doctor example: $P = 0.057$ $(\Phi(-1.58))$.
- Chocolatier example: $P = 0.0025$ $(\Phi(-2.8))$.
- Car example: $P = 0.0734$ $(\Phi(1.79) = 0.9633)$.

## 16.4 Practice problems

**Problem 16.1.**
A researcher is interested in whether free-range chickens produce more nutritious eggs than caged chickens, and decides to test this by measuring the amount of DHA, a type of omega-3 fatty acid, in eggs. First the researcher looks up the average amount of DHA in eggs from conventional, caged chickens and learns this is 90mg per egg on average. The researcher then takes a sample of 50 eggs from free-range chickens and measures the amount of DHA in each egg. From this the researcher found a sample average of 95mg of DHA in these eggs, with a sample standard deviation of 17mg of DHA.

(a) If the null hypothesis is that free-range chicken eggs are nutritionally equivalent to those of caged eggs (as measured by the amount of DHA in the eggs) and the alternative hypothesis is that free-range chicken eggs are more nutritious, should the researcher reject the null hypothesis with a significance level of $\alpha = 0.05$?

(b) Should the researcher reject the null hypothesis if the significance level is $\alpha = 0.01$?

**Problem 16.2.**
An engineer at a company that produces batteries for smart phones believes she can improve the battery life of phones by a simple modification of the battery's design. The company decides to test the engineer's modified design by creating a sample batch of sixteen batteries with this design. The company knows that the battery life of a phone between charges with the old design is normally distributed with mean 8 hours and standard deviation 15 minutes. If the sample mean of the lifetime for the modified battery design averaged 8 hours and 15 minutes, is there enough evidence to conclude at the $\alpha = 0.01$ significance level that batteries with the new, modified design last longer than the batteries with the old design? (Use $z_{0.01} = 2.33$.)

# Part V

# Appendices

# A

# *Integration in Multiple Variables*

In order to do computations with joint probability distributions, we will need to know how to integrate functions in several variables. This set of notes is meant to be a quick introduction for students that have not taken multivariable calculus, or a refresher for students that need a review.

## A.1 Review of Integration in One Variable

Recall that if $f : [a, b] \to \mathbb{R}$ is a continuous function, the ***(definite) integral*** of $f$ is defined as a limit of Riemann sums. In particular, we choose a partition $\mathcal{P}$ of $[a, b]$:

$$\mathcal{P} = \{x_0, x_1, x_2, ..., x_n\},$$
$$\text{where } a = x_0 < x_1 < x_2 < ... < x_n = b.$$

A ***Riemann sum*** of $f$ with respect to the partition $\mathcal{P}$ is the quantity

$$\sum_{i=1}^{n} f(x_i^*)\Delta x_i,$$

where $\Delta x_i = x_i - x_{i-1}$ (this is the length of the $i$-th subinterval in the partition) and $x_i^*$ is any point in $[x_{i-1}, x_i]$. Obviously the value of this sum depends on the choice of $\mathcal{P}$ and the choice of each $x_i^*$. Incredibly, if we take the limit as the pieces of the partition get arbitrarily small, we always get the same value, regardless of the $\mathcal{P}$ and $x_i^*$'s we choose in calculating each of the Riemann sums.

Writing $|\mathcal{P}| = \max_i \Delta x_i$ (so $|\mathcal{P}|$ is the length of the widest subinterval determined by $\mathcal{P}$) the integral of $f$ over $[a, b]$ is defined as

$$\int_a^b f(x)\,dx = \lim_{|\mathcal{P}|\to 0} \sum_{i=1}^{n_\mathcal{P}} f(x_i^*)\Delta x_i.$$

The number of terms in the sum depends on the partition $\mathcal{P}$ we choose. Here we're letting $n_\mathcal{P}$ denote the number of subintervals into which $[a, b]$ is partitioned into by $\mathcal{P}$.

Since we're taking a limit, we always have to ask ourselves if this limit exists or not. It is a theorem (that we won't try to prove) that this limit will always exist if $f$ is continuous and $[a, b]$ is a closed, bounded interval.

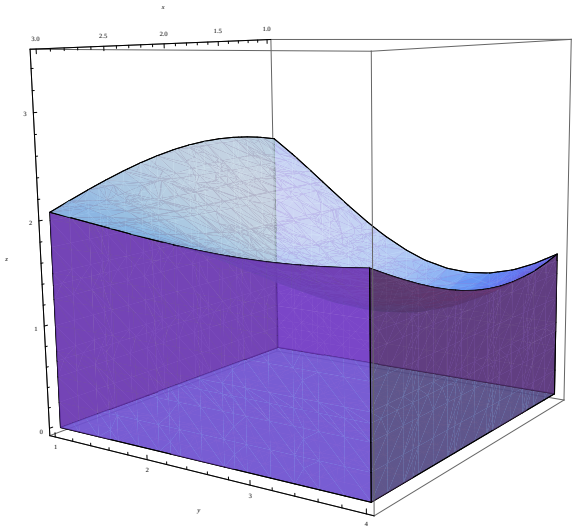
When you first learn about integration, you build the integral up for a singular purpose: to find the area between some curve $y = f(x)$ and the $x$-axis. The idea is to approximate the area under the curve with things that

are much simpler to work with: rectangles. In the Riemann sum, the value $f(x_i^*)$ acts as the height of the $i$-th rectangle, and the $\Delta x_i$ is the width. So we calculate the area of each rectangle and add them all up.

Of course, you realize very quickly that integrals can do much more than simply calculate areas. Integration is ubiquitous in mathematics: from geometry to statistics to physics, integrals are everywhere. The reason that integrals are such a useful tool is that they can be thought of as a very special type of infinite summation. The integral $\int_a^b f(x)\,dx$ is, in some sense, the result of summing up the values of $f(x)$ for every single $x$ in $[a, b]$; it's just that we weight the values in the sum by a very small number (this is basically what the $dx$ is) to keep this "sum" from blowing up to infinity.

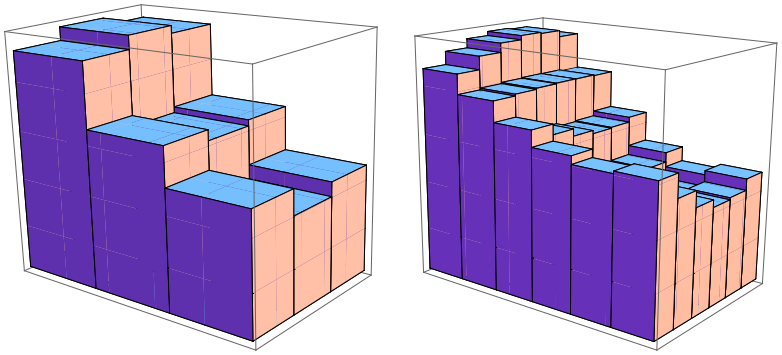
## Calculating Volumes

To motivate integration in several variables, consider the following problem. Suppose that $f : D \to \mathbb{R}$ is a continuous function and that $D$, the domain of $f$, is a rectangle in $\mathbb{R}^2$. Suppose also that $f(x, y) \geq 0$ for all $(x, y) \in D$. We now construct a three-dimensional object whose top is $z = f(x, y)$, whose bottom is $D$, and we fill in all of the space in-between. See the figure below.



Now we want to determine what the volume of this solid is. To do this we do the same sort of thing we did to calculate the area under a curve: approximate the volume with simpler objects. The simpler objects we'll use are rectangular prisms. If a prism has height $h$, length $\ell$, and width $w$, then we know its volume is $h\ell w$.

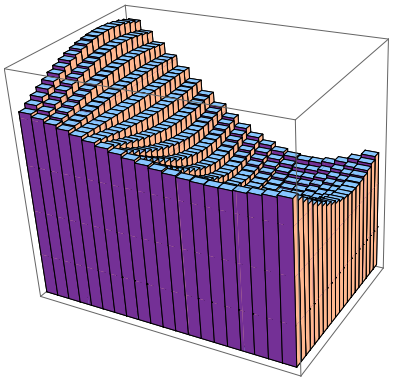
So what we'll do is cram several rectangular prisms under the surface $z = f(x, y)$, determine the volume of each prism, and then finally sum up these volumes. See Figure A.1.



(a) 9 prisms



(b) 36 prisms



(c) 400 prisms

Figure A.1: Approximating volume with prisms.

Suppose we label these prisms $P_1, P_2, ..., P_n$ and let $h_i$, $\ell_i$, and $w_i$ denote the height, length, and width of each $P_i$. Then we know that the volume of our object is approximated by

$$\text{Volume} \approx \sum_{i=1}^{n} h_i w_i \ell_i.$$

Of course what we want to do is take the limit as we fill the area under the curve with skinnier and skinnier prisms. In order to do this we need to state precisely how these these prisms are placed beneath the surface.

Suppose the four corners of the domain $D$ are $(a, c), (b, c), (b, d), (a, d)$. See Figure A.2. This means that the rectangle $D$ consists precisely of those

points $(x, y)$ where $a \leq x \leq b$ and $c \leq y \leq d$. So we can express $D$ as the set

$$D = [a, b] \times [c, d] = \left\{ (x, y) \,\middle|\, a \leq x \leq b,\, c \leq y \leq d \right\}.$$

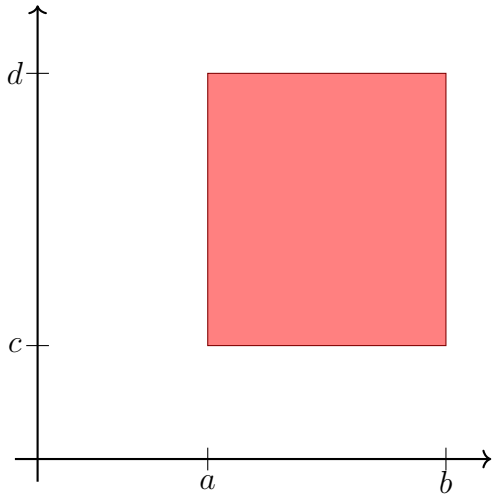



Figure A.2: The rectangle $D$ is the domain of our function.

Our first step is to cut $D$ into finitely many rectangular pieces; each piece will serve as the base of a rectangular prism. To do this we'll cut $[a, b]$ into subintervals by using the partition $\mathcal{P} = \{x_0, x_1, ..., x_m\}$ where

$$a = x_0 < x_1 < ... < x_m = b,$$

and we'll cut $[c, d]$ into subintervals with the partition $\mathcal{Q} = \{y_0, y_1, ..., y_n\}$ where

$$c = y_0 < y_1 < ... < y_n = d.$$

This partitions $D$ into $mn$ subrectangles. We'll let the rectangle in the $i$-th column and $j$-th row (ordered left-to-right, bottom-to-top) be denoted $D_{ij}$. See Figure A.3.

Let's denote the area of the rectangle $D_{ij}$ by $\Delta A_{ij}$. (Notice $\Delta A_{ij} = \Delta x_i \cdot \Delta y_j$.)

Now that we have bases for our rectangular prisms, we just need to determine their height. To do this we let $P_{ij}^*$ denote any point inside of $D_{ij}$, and then use $f(P_{ij}^*)$ as the height of the prism. Notice that since the $x$-coordinates of $P_{ij}^*$ are in the $i$-th subinterval of $[a, b]$, and the $y$-coordinates of $P_{ij}^*$ are in the $j$-th subinterval of $[c, d]$, we have $P_{ij}^* = (x_i^*, y_j^*)$. Thus the volume of the $ij$-th prism is

$$f(x_i^*, y_j^*)\Delta A_{ij}.$$

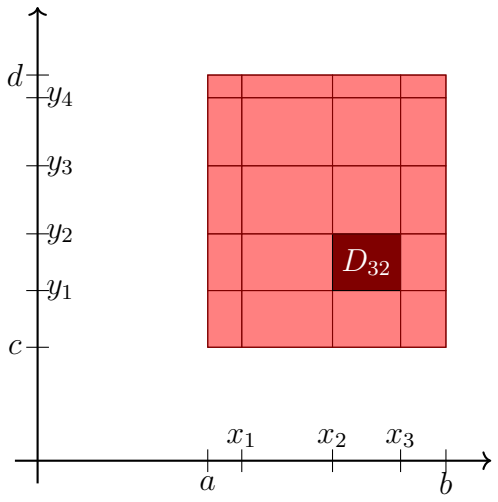



Figure A.3: The rectangle $D$ is partitioned into subrectangles.

Summing up the volumes of each of these prisms, we have an estimate for the volume of our solid:

$$\text{Volume} \approx \sum_{i=1}^{m} \sum_{j=1}^{n} f(x_i^*, y_j^*) \Delta A_{ij}.$$

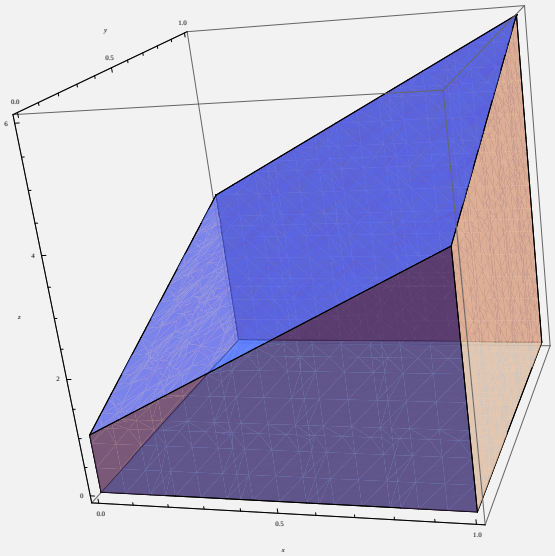
To get a better approximation, stick more, skinnier, prisms underneath the surface. To get the "best" approximation (i.e., the true volume), take the limit as the prisms get arbitrarily skinny. To do this we need that both the widths and lengths of our base rectangles get arbitrarily small. That is, we require $|\mathcal{P}| \to 0$ and $|\mathcal{Q}| \to 0$.

The limit as the rectangles get arbitrarily skinny is called the ***(double) integral of $f$ over the rectangle $D$*** and is denoted as follows:

$$\iint_D f(x,y)\, dA = \lim_{|\mathcal{P}|\to 0} \lim_{|\mathcal{Q}|\to 0} \sum_{i=1}^{m_{\mathcal{Q}}} \sum_{j=1}^{n_{\mathcal{P}}} f(x_i^*, y_j^*)\, \Delta A_{ij}.$$

As always, we have to worry about whether this limit exists or not. As in the case of one variable, there is a theorem that says that this limit will exist as long as our function $f$ is continuous and $D$ is a rectangle of finite area. (There are other, more general, conditions which guarantee the integral exists, but this is good enough for right now.)

**Example A.1.**
Calculate the volume between the surface $z = 3x + 2y + 1$ and the $xy$-plane over the unit square, $D = [0, 1] \times [0, 1]$.



To make this process as easy as possible, let's say that we partition both the the horizontal and vertical intervals $[0, 1]$ into $n$ subintervals of equal width, and use the upper, right-hand corner of each rectangle as the point where we'll evaluate the function to determine the height of a prism. (As long as our function is continuous we'll get the same value in the end, so we can pick points that are easy to work with.)

This gives us $n^2$ subrectangles, each of area $1/n^2$, and $x_i^* = 1/n$, $y_j^* = 1/n$. Thus our volume is given by the limit:

$$\begin{aligned}\iint_{[0,1]\times[0,1]} (3x + 2y + 1)\, dA &= \lim_{n\to\infty} \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} (3x_i^* + 2y_j^* + 1) \cdot \frac{1}{n^2} \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{3i}{n} + \frac{2j}{n} + 1 \right) \cdot \frac{1}{n^2}\end{aligned}$$

Notice that since we cut both the horizontal and vertical intervals into $n$ pieces we have $m = n$, which is why we have two limits as $n \to \infty$ on the first line. Of course, taking the first (inner) limit gives us a number, and so taking the second (outer) limit doesn't do anything,

so we can drop one of the limits.

$$\begin{aligned}
\iint_{[0,1]\times[0,1]} (3x+2y+1)\,dA &= \lim_{n\to\infty} \sum_{i=1}^{n}\sum_{j=1}^{n} \left(\frac{3i}{n^3}+\frac{2j}{n^3}+\frac{1}{n^2}\right) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \left(\sum_{j=1}^{n}\frac{3i}{n^3}+\sum_{j=1}^{n}\frac{2j}{n^3}+\sum_{j=1}^{n}\frac{1}{n^2}\right) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{3i}{n^2}+\frac{1}{n}+\frac{2}{n^3}\sum_{j=1}^{n} j\right) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \left(\frac{3i}{n^2}+\frac{1}{n}+\frac{2}{n^3}\cdot\frac{n^2+n}{2}\right) \\
&= \lim_{n\to\infty} \left(\sum_{i=1}^{n}\frac{3i}{n^2}+\sum_{i=1}^{n}\frac{1}{n}+\sum_{i=1}^{n}\frac{n^2+n}{n^3}\right) \\
&= \lim_{n\to\infty} \left(\frac{3}{n^2}\sum_{i=1}^{n} i+1+\frac{n^2+n}{n^2}\right) \\
&= \lim_{n\to\infty} \left(\frac{3}{n^2}\cdot\frac{n^2+n}{2}+1+\frac{n^2}{n^2}+\frac{n}{n^2}\right) \\
&= \lim_{n\to\infty} \left(\frac{3}{2}\cdot\frac{n^2+n}{n^2}+2+\frac{1}{n}\right) \\
&= \lim_{n\to\infty} \left(2+\frac{3}{2}+\frac{3}{2n}+\frac{1}{n}\right) \\
&= \frac{7}{2}
\end{aligned}$$

Evaluating the limit above we used two facts that you learned in single variable calculus, but may have forgotten about: you can "distribute" summations:

$$\sum_i (a_i+b_i) = \sum_i a_i + \sum_i b_i,$$

and there's a nice formula for the sum $1+2+3+\cdots+n$:

$$\sum_{i=1}^{n} i = \frac{n^2+n}{2}.$$

So the volume between the plane $3x+2y+1$ and the unit square in the $xy$-plane is $7/2$.

Of course, evaluating a limit such as the one above is a rather tedious thing to do. It'd be nice we if we had some way to turn these complicated double integrals into "normal" integrals of one variable where we could use tools such as integration by parts and $u$-substitutions. We'll see how this can be done later. For the time being we'll just be content with the fact that we can, at least in principle, evaluate these double integrals by taking limits.

## Properties of the Integral

The double integral satisfies several properties analogous to properties of integrals of single variables. Here we mention a few of the most basic ones.

(i) Letting $\text{Area}(D)$ denote the area of a rectangle $D$,

$$\iint_D 1 \, dA = \text{Area}(D).$$

This is straight forward to see: the double sum we're taking the limit of is just

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \Delta A_{ij},$$

where $\Delta A_{ij}$ is the area of the $ij$-th subrectangle. However we're summing this over all of the subrectangles, so we just get back the area of $D$.

Note: Notationally, we sometimes write

$$\iint_D dA = \iint_D 1 \, dA.$$

(ii) If $\lambda \in \mathbb{R}$ is a constant and $f : D \to \mathbb{R}$ is continuous, then

$$\iint_D \lambda f(x, y) \, dA = \lambda \iint_D f(x, y) \, dA.$$

This follows from the fact that we can pull the constant $\lambda$ out of the sums and limits in the definition of the integral.

(iii) If $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ are both continuous, then

$$\iint_D (f(x, y) + g(x, y)) \, dA = \iint_D f(x, y) \, dA + \iint_D g(x, y) \, dA.$$

This again just follows from the fact that we can split sums and limits up across addition. Notice that this, combined with the previous property, means that we can also split up a subtraction: write $f(x,y) - g(x,y) = f(x,y) + (-1) \cdot g(x,y)$.

(iv) If $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ are both continuous and $f(x,y) \leq g(x,y)$ for every $(x,y) \in D$, then

$$\iint_D f(x,y)\,dA \leq \iint_D g(x,y)\,dA.$$

Once more, this follows by the analogous properties for sums and limits.

## Applications

Even though we built the double integral above for the purpose of calculating volumes, it's clear that the definition still "makes sense" for functions which may be negative. If the function is negative – i.e., if the surface $z = f(x,y)$ is below the $xy$-plane – then the corresponding integral we calculate will be negative. This is similar to how $\int_a^b f(x)\,dx < 0$ if $y = f(x)$ stays below the $x$-axis. In general a function may be above the $xy$-plane sometimes, and below the $xy$-plane at other times. When this happens, the integral brakes up into we have positive and negative pieces, and may cancel out. In these cases the double integral doesn't represent a volume, but may still have a concrete, physical meanings.

After we've developed some more tools for calculating integrals, we'll consider the more applications in detail, but it's worthwhile to go ahead and mention some of the things these double integrals can be used for:

**Mass of an object**

If we have a rectangular "sheet" of some material (metal, plastic, cloth, ...), and if we know the know what the density of this material is at any point, integrating the density gives us the mass of the object. Say our rectangular "sheet" is $w \times \ell$. We can think of this as the rectangle $[0,w] \times [0,\ell]$ in the $xy$-plane. For any point $(x,y)$ inside the rectangle, suppose that $\rho(x,y)$ represents the density of the material at that particular point. Then the mass of the sheet is

$$m = \iint_{[0,w]\times[0,\ell]} \rho(x,y)\,dA.$$

If you consider how density is defined in physics, this is almost obvious. Density, in two dimensions, is mass divided by area: $\rho = \frac{m}{A}$. So over a very small subrectangle, $D_{ij}$, the density is approximately

$$\rho_{D_{ij}} \approx \frac{\text{mass of } D_{ij}}{\text{area of } D_{ij}} = \frac{\text{mass of } D_{ij}}{\Delta A_{ij}}.$$

When we write out the limit of Riemann sums, the $\Delta A_{ij}$'s cancel out and we're just summing up the mass of little pieces of $D$.

**Average value**

If $f : D \to \mathbb{R}$ is a continuous function on a rectangle $D$, then there may be times we want to know what the average value of $f$ is. For example, suppose that $D$ represents the floor in a room, and for each point in the room, the temperature you record at that point is determined by where you're standing in the room – by your $xy$-coordinates on the floor. (Of course, the temperature in a real room may also depend on how high above the floor you are.) If $T(x, y)$ gives the temperature over the point $(x, y)$, then the average temperature in the room is

$$\text{Avg. temp} = \frac{1}{\text{Area(D)}} \iint_D T(x, y)\, dA.$$

Why is this the average temperature? If the temperature throughout the entire room was constant, say $T(x, y) = C$, then we'd say that the average temperature in the room was $C$. So to estimate the average temperature, let's partition the room into very small rectangles and suppose that the temperature is constant on each of those rectangles (possibly a different constant on different rectangles).

Now if we wanted to combine all of these average temperatures over small regions together to get the average temperature of the whole region, we'd have to weight those averages by the relative size of the region; that is, by how much proportion of the room is taken up by that region. (Why? Because a 1-inch $\times$ 1-inch region where the temperature is $90°F$ doesn't contribute as much to the average as a 10-ft $\times$ 10-ft region where the temperature is $90°F$. If the temperature is really warm over a large region, that counts a lot more for the average than being really warm over a very small region.)

Let's suppose that we call the subrectangles of our partition $D_{ij}$, with area $\Delta A_{ij}$. The proportion of the room taken up by $D_{ij}$ is $\frac{\Delta A_{ij}}{\text{Area}(D)}$. Say the temperature we use for the constant on $D_{ij}$ is $T(x_i^*, y_j^*)$. So

we sum up the values

$$\sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i^*, y_j^*)\frac{\Delta A_{ij}}{\text{Area}(D)}$$

$$=\frac{1}{\text{Area}(D)}\sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i^*, y_j^*)\Delta A_{ij}.$$

Taking the limit gives exactly the integral described above. Notice that it makes perfect sense to talk about an average temperature being negative!

Of course, there's nothing special about the fact that we're talking about temperature above. In general, the ***average value*** of a continuous function over a rectangle $D$ is

$$\text{Average of } f = \frac{1}{\text{Area}(D)}\iint_D f(x,y)\,dA.$$

## A.2 Iterated Integrals

We now turn our attention to how to evaluate the integrals defined above *without* having to compute a limit of Riemann sums

### Motvation & "Partial Integration"

We defined the double integral of a continuous function as a limit of a double Riemann sum above. While this definition makes intuitive sense (approximating a quantity with simpler quantities and taking a limit to get the "best" approximation), it's typically extremely difficult and tedious to use for calculations. Now we want to introduce a way of calculating these quantities which will allow us to apply the tools and techniques from integration in one variable. Before we describe how this is done, we need to make one technical detour.

Recall that the partial derivatives, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$, are calculated by keeping one variable constant and differentiating with respect to the other variable. Suppose we instead want a "partial antiderivative" of a function. That is, suppose that $f(x,y)$ is a given function. Can we find functions $G(x,y)$ and $H(x,y)$ such that $\frac{\partial G}{\partial x} = f(x,y)$ and $\frac{\partial H}{\partial y} = f(x,y)$? If we were considering functions of a single variable, then we'd just integrate the function to get

its antiderivative. Since partial derivatives are calculated by keeping one variable constant, these "partial antiderivatives" can be calculated the same way: integrate the function by keeping one variable constant.

That is, to calculate $G(x, y)$ we'll integrate $f(x, y)$ with respect to $x$, pretending that the $y$ in our function is a constant. Similarly, to calculate $H(x, y)$ we integrate $f(x, y)$ with respect to $y$, pretending $x$ is constant. This is denoted as follows:

$$G(x, y) = \int f(x, y)\, dx$$
$$H(x, y) = \int f(x, y)\, dy$$

There is one caveat here: when we calculate $\int f(x, y)\, dx$ instead of picking up a $+C$, we pick up a $+k(y)$. That is, since $y$'s are constant when we calculate $\frac{\partial G}{\partial x}$, any function of $y$ is also constant. So our $+C$ can be any expression that involves only $y$'s: from the partial derivative point of view these are functions. Similarly, when we calculate $\int f(x, y)\, dy$, we pick up a $+\ell(x)$.

**Example A.2.**
Find a $G(x, y)$ such that $\frac{\partial G}{\partial x} = x^2y - \sin(xy)$. Find a $H(x, y)$ such that $\frac{\partial H}{\partial y} = x^2y - \sin(xy)$.

We simply integrate, pretending one variable or the other is a constant.

$$\begin{aligned} G(x, y) &= \int \left(x^2y - \sin(xy)\right)\, dx \\ &= \frac{x^3y}{3} + \frac{\cos(xy)}{y} + k(y) \end{aligned}$$

$$\begin{aligned} H(x, y) &= \int \left(x^2y - \sin(xy)\right)\, dy \\ &= \frac{x^2y^2}{2} + \frac{\cos(xy)}{x} + \ell(x) \end{aligned}$$

Now let's just double check that these are the functions we want:

$$\begin{aligned}\frac{\partial G}{\partial x} =& \frac{\partial}{\partial x}\left(\frac{x^3 y}{3} + \frac{\cos(xy)}{y} + k(y)\right)\\ =& \frac{3x^2 y}{3} + \frac{-\sin(xy)\, y}{y} + 0\\ =& x^2 y - \sin(xy)\end{aligned}$$

$$\begin{aligned}\frac{\partial H}{\partial y} =& \frac{\partial}{\partial y}\left(\frac{x^2 y^2}{2} + \frac{\cos(xy)}{x} + \ell(x)\right)\\ =& \frac{2x^2 y}{2} + \frac{-\sin(xy)\, x}{x} + 0\\ =& x^2 y - \sin(xy)\end{aligned}$$

## Iterated Integrals

To calculate a double integral,

$$\iint\limits_{[a,b]\times[c,d]} f(x,y)\, dA,$$

we will convert the double integrals into two integrals of a single variable, combined together in a particular way. The basic idea is the following: Geometrically, double integrals were developed for calculating volumes. However these is another way to calculate volumes, provided that you know the cross-sectional areas of the solid you're integrating.

Recall that if we have a solid positioned in three-dimensional space so that the $x$-axis runs through the solid, like a chicken on a rotisserie, then for each plane $x = c$ we denote the area of the intersection of the plane and the solid by $A(x)$. Then the volume of the solid is given by integrating $A(x)$:

$$\text{Volume} = \int_a^b A(x)\, dx$$

For example, above we considered the volume of the solid whose top was the plane $3x + 2y + 1$, and whose bottom was the unit square $[0,1] \times [0,1]$. Cutting this surface with a plane we see the blue region plotted in Figure A.4.

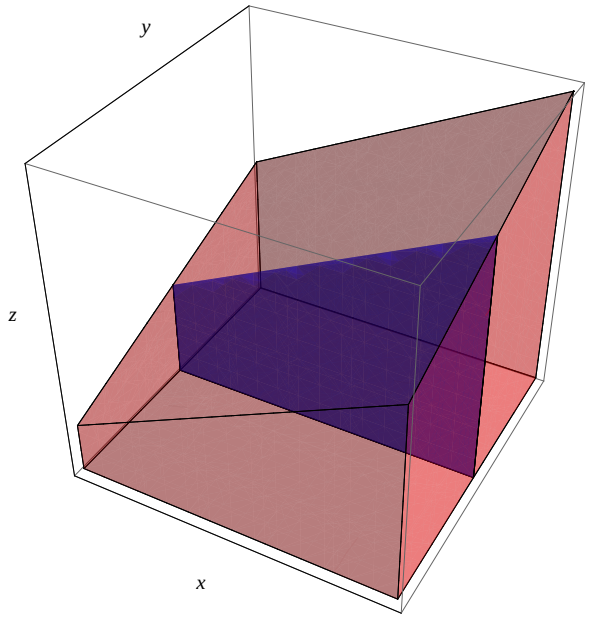



Figure A.4: Cutting a solid with a plane.

If we could calculate the area, $A(x)$, of this blue slice of the surface, we'd integrate $\int_0^1 A(x)\,dx$ to find the volume. There's nothing special about the $x$ axis here: we could just as easily consider slices of the surface given by planes $y = c$, call $A(y)$ the area of these slices, and then integrate $\int_0^1 A(y)\,dy$ to get the volume.

Calculating these cross-sectional areas is actually very easy because they're just the area under the curve. In particular, the area $A(x)$ is given by

$$A(x) = \int_0^1 (3x + 2y + 1)\ dy.$$

This is just the area of the blue slice because the blue slice is the area underneath the curve $3x + 2y + 1$. Here we've set $x$ to be a constant, so $y$ is the only quantity that changes. Performing the integration we see that this really is just a function of $x$: the $y$'s get replaced with numbers when we do the integration.

$$\begin{aligned} A(x) &= \int_0^1 (3x + 2y + 1)\ dy \\ &= \left(3xy + y^2 + y\right)\Big|_0^1 \\ &= 3x + 2. \end{aligned}$$

This is the cross-sectional area of our blue slice. Integrating this quantity we get the volume.

$$\begin{aligned}\text{Volume} &= \int_0^1 A(x)\,dx \\ &= \int_0^1 (3x+2)\ dx \\ &= \left(\frac{3x^2}{2}+2x\right)\Bigg|_0^1 \\ &= \frac{3}{2}+2 \\ &= 7/2\end{aligned}$$

Usually we don't bother to write down $A(x)$ as a separate function, and instead just plug our expression for $A(x)$,

$$\int_0^1 (3x+2y+1)\ dy,$$

into the integral:

$$\text{Volume} = \int_0^1 \int_0^1 (3x+2y+1)\ dy\,dx.$$

To evaluate an expression like this we work "inside-out," starting with the inner-most integral and integrating piece by piece until we've evaluated all of the integrals.

$$\begin{aligned}\text{Volume} &= \int_0^1 \int_0^1 (3x+2y+1)\ dy\,dx \\ &= \int_0^1 \left(3xy+y^2+1\right)\Big|_0^1\ dx \\ &= \int_0^1 (3x+2)\ dx \\ &= \left(\frac{3x^2}{2}+2x\right)\Bigg|_0^1 \\ &= \frac{3}{2}+2 \\ &= 7/2\end{aligned}$$

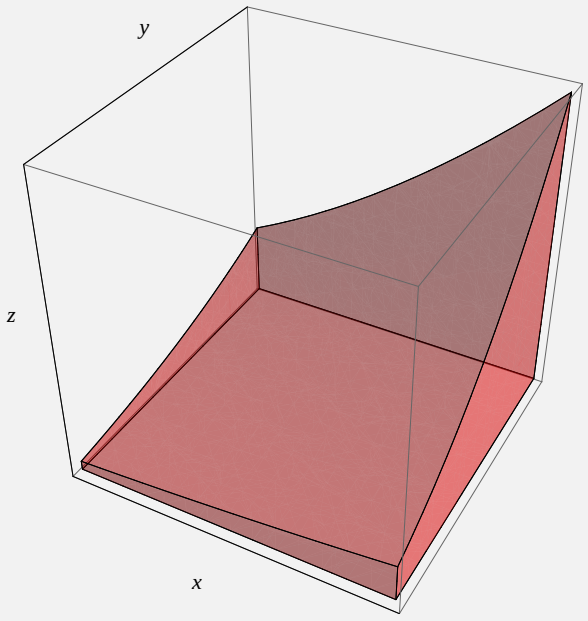
The procedure outlined above is generalized by the following theorem.

**Theorem A.1** (Fubini's theorem)**.**
*If $f(x, y)$ is a continuous function defined on the rectangle $D = [a, b] \times [c, d]$, then*

$$\iint_D f(x,y)\,dA = \int_a^b \int_c^d f(x,y)\,dy\,dx = \int_c^d \int_a^b f(x,y)\,dx\,dy.$$

**Example A.3.**
Calculate the integral

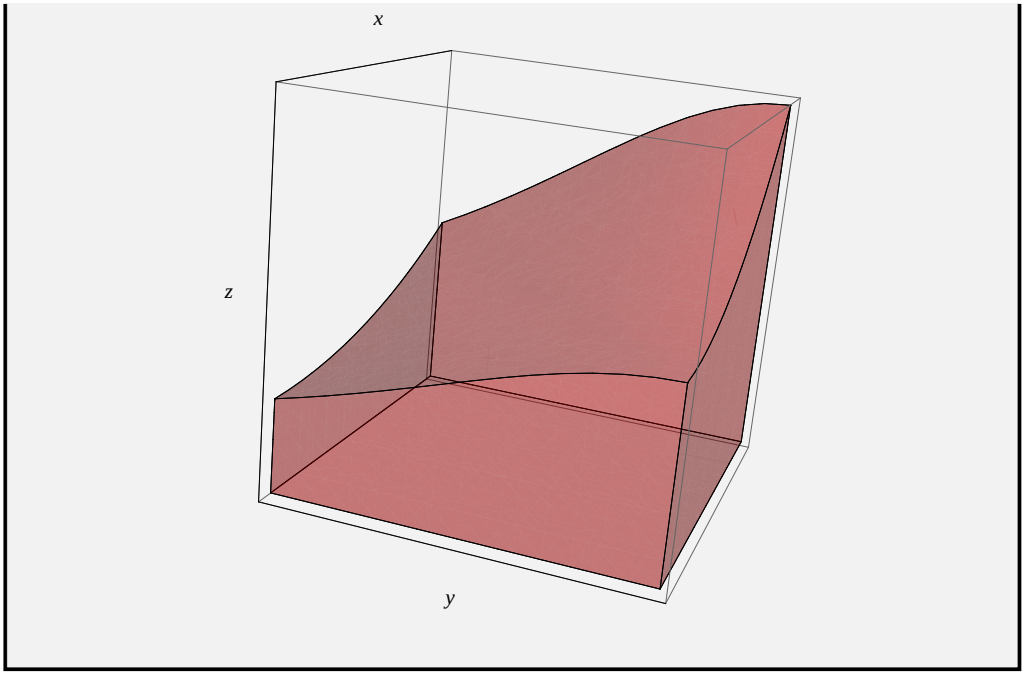
$$\iint_{[2,4]\times[1,2]} \frac{x^2\,y^3}{2}\,dA.$$



$$\begin{aligned}
\iint_{[2,4]\times[1,2]} \frac{x^2\,y^3}{2}\,dA =& \frac{1}{2}\iint_{[2,4]\times[1,2]} x^2\,y^3\,dA \\
=& \frac{1}{2}\int_2^4\int_1^2 x^2\,y^3\,dy\,dx \\
=& \frac{1}{2}\int_2^4 \frac{x^2\,y^4}{4}\bigg|_1^2\,dx \\
=& \frac{1}{2}\int_2^4 \left(4x^2 - \frac{x^2}{4}\right)\,dx \\
=& \frac{1}{2}\left(\frac{4x^3}{3} - \frac{x^3}{12}\right)\bigg|_2^4 \\
=& \frac{1}{2}\left(\frac{4\cdot 64}{3} - \frac{64}{12} - \frac{8\cdot 4}{3} + \frac{8}{12}\right) \\
=& \frac{1}{2}\left(\frac{256}{3} - \frac{16}{3} - \frac{32}{3} + \frac{2}{3}\right) \\
=& \frac{210}{6} \\
=& \frac{105}{3} \\
=& 35
\end{aligned}$$

**Example A.4.**
Calculate the integral

$$\iint_{[0,1]\times[0,1]} \frac{1+x^2}{1+y2}\,dA$$




$$\iint_{[0,1]\times[0,1]} \frac{1+x^2}{1+y^2}\,dA = \int_0^1 \int_0^1 \frac{1+x^2}{1+y^2}\,dx\,dy$$

Notice that, with respect to $x$, $\frac{1}{1+y^2}$ is a constant. Hence we can pull it out of the inner-most integral:

$$\int_0^1 \int_0^1 \frac{1+x^2}{1+y^2}\,dx\,dy = \int_0^1 \frac{1}{1+y^2} \int_0^1 \left(1+x^2\right)\,dx\,dy$$

Now once we integrate, the value $\int_0^1 \left(1+x^2\right)\,dx$ is just a number, so we can pull it out of the outer-most integral:

$$\begin{aligned}
\int_0^1 \frac{1}{1+y^2} \int_0^1 \left(1+x^2\right)\,dx\,dy &= \int_0^1 \left(1+x^2\right)\,dx \cdot \int_0^1 \frac{1}{1+y^2}\,dy \\
&= \left(x+\frac{x^3}{3}\right)\Big|_0^1 \cdot \tan^{-1}(y)\Big|_0^1 \\
&= \frac{4}{3}\cdot\frac{\pi}{4} \\
&= \pi/3
\end{aligned}$$

# A.3 Double Integrals Over General Regions

## Motivating Example

A silicon wafer is a large, circular disc made of silicon which is used in the manufacture of computer processors and other electronic devices. Suppose that in the process of fabricating such a wafer some impurities are introduced (dust, water vapor, etc.) so that the wafer isn't pure silicon. If we were able to determine precisely where these impurities lie in the wafer, then we might be able to determine the density of the wafer at a particular point. To figure out the mass of the entire wafer we could then integrate this density. This presents us with a problem in that the wafer is circular (so the domain of our density function is a disc in the plane), whereas we only know how to integrate functions with a rectangular domain. So we need some way of extending our usual double integrals to deal with functions with other sorts of domains.

To associate some actual numbers with the scenario described above, suppose that our wafer has a radius of one meter, and for a point $(x, y)$ in the wafer, the density of the wafer at that point is

$$\rho(x, y) = x^2 \cos(y) + 1.$$

The domain of our function $\rho$ is

$$D = \left\{ (x, y) \middle| x^2 + y^2 \leq 1 \right\}$$

Let's notice that if we pick an $x$-coordinate of a point inside this disc, the $y$-coordinates we can tack onto this $x$-coordinate lie between the values $-\sqrt{1 - y^2}$ and $\sqrt{1 - y^2}$. So for example, if we look at all of the $(x, y)$ points inside our disc where the $x$-coordinate is $1/2$, the $y$-coordinates have to be between $-\sqrt{1 - 1/4}$ and $\sqrt{1 - 1/4}$.

Recall from above that to evaluate the integral

$$\iint_D \rho(x, y)\, dA,$$

we integrate a "cross-section" function, $A(x)$. If we knew what the cross-section was, then we'd integrate

$$\iint_D \rho(x, y)\, dA = \int_{-1}^{1} A(x)\, dx,$$

since our $x$'s run from $-1$ to 1. To calculate the cross-section function last time we just integrated our initial function $\rho(x, y)$ with respect to $y$ over all of the possible $y$-values. Here our $y$-values depend on our chosen $x$, but once we've chosen an $x$ we expect that our cross-section function should be

$$A(x) = \int_{-\sqrt{3}/2}^{\sqrt{3}/2} \rho(x, y)\, dy = \int_{-\sqrt{3}/2}^{\sqrt{3}/2} (x^2 \cos(y) + 1)\, dy.$$

Of course, there's nothing special about the choice of $x = 1/2$. In general, for any $x$ between $-1$ and 1, the cross-section is

$$\begin{aligned} A(x) =& \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (x^2 \cos(y) + 1)\, dy \\ =& (x^2 \sin(y) + y)\Big|_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \\ =& x^2 \sin(\sqrt{1-x^2}) + \sqrt{1-x^2} - x^2 \sin(-\sqrt{1-x^2}) - \sqrt{1-x^2} \\ =& 2x^2 \sin(\sqrt{1-x^2}) \end{aligned}$$

Above we used the fact that $\sin\theta$ is an odd function: $\sin(-\theta) = -\sin\theta$.

Notice that in order to find this cross-section, the bounds of our integral had to depend on where we were trying to find the cross-section. Aside from this one modification, our cross-section was found exactly like before. Notice here the bounds for our integral with respect to $y$ were functions of $x$.

Now that we have the cross-section, we can calculate the integral we initially wanted:

$$\begin{aligned} \iint_D \rho(x, y)\, dA &= \int_{-1}^{1} A(x)\, dx \\ &= \int_{-1}^{1} 2x^2 \sin(\sqrt{1-x^2})\, dx \end{aligned}$$

This is a hard integral to solve, so we won't bother to explicitly solve it right now, but just content ourselves with the fact that we can rewrite the integral over a non-rectangular region as an iterated integral:

$$\iint_D (x^2 \cos(y) + 1)\, dA = \int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} (x^2 \cos(y) + 1)\, dy\, dx.$$

## Integrals Over General Regions

We'll say a subset $D$ of the plane $\mathbb{R}^2$ is ***type I*** if the $x$-values of points in $D$ stay inside some fixed interval $[a, b]$, but the $y$-values of points in $D$ are bounded by functions of $x$. That is, a type I region can be written as

$$D = \left\{ (x, y) \middle| a \le x \le b,\ g(x) \le y \le h(x) \right\}.$$

The integral of a continuous function $f(x, y)$ over a type I region $D$ is given by

$$\iint_D f(x, y)\, dA = \int_a^b \int_{g(x)}^{h(x)} f(x, y)\, dy\, dx.$$

We say that $D$ is ***type II*** if the roles of $x$ and $y$ are switched from that of a type I region: that is, a type II region $D$ can be written as

$$D = \left\{ (x, y) \middle| c \le y \le d,\ k(y) \le x \le \ell(x) \right\}.$$

The integral of a continuous $f(x, y)$ over a type II region $D$ is

$$\iint_D f(x, y)\, dA = \int_c^d \int_{k(y)}^{\ell(y)} f(x, y)\, dx\, dy.$$

(Notice that some regions are both type I and type II simultaneously. For example, the disc considered above could be considered as type I or type II.)

**Example A.5.**
Evaluate the integral

$$\iint_D e^{x/y}\, dA$$

where $D$ is the region

$$D = \left\{ (x, y) \middle| 1 \le y \le 2,\ y \le x \le y^3 \right\}.$$

Notice that this is a type II region, so we our integral is given by

$$\begin{aligned}
\iint_D e^{x/y}\,dA &= \int_1^2 \int_y^{y^3} e^{x/y}\,dx\,dy \\
&= \int_1^2 \int_y^{y^3} e^{x\cdot 1/y}\,dx\,dy \\
&= \int_1^2 \left(\frac{e^{x/y}}{1/y}\right)\Bigg|_y^{y^3} dy \\
&= \int_1^2 ye^{x/y}\Big|_y^{y^3} dy \\
&= \int_1^2 \left(ye^{y^2} - ye\right)\,dy \\
&= \int_1^2 ye^{y^2}\,dy - \int_1^2 ye\,dy
\end{aligned}$$

For the integral on the left, perform the substitution $u = y^2$, $du = 2ydy$.

$$\begin{aligned}
\int_1^2 ye^{y^2}\,dy - \int_1^2 ye\,dy =& \frac{1}{2}\int_1^4 e^u\,du - \int_1^2 ye\,dy \\
=& \frac{1}{2}e^u\Big|_1^4 - \frac{ey^2}{2}\Big|_1^2 \\
=& \frac{1}{2}\left(e^4 - e\right) - \left(\frac{4e}{2} - \frac{e}{2}\right) \\
=& \frac{1}{2}(e^4 - e - 3e) \\
=& \frac{e^4 - 4e}{2}
\end{aligned}$$

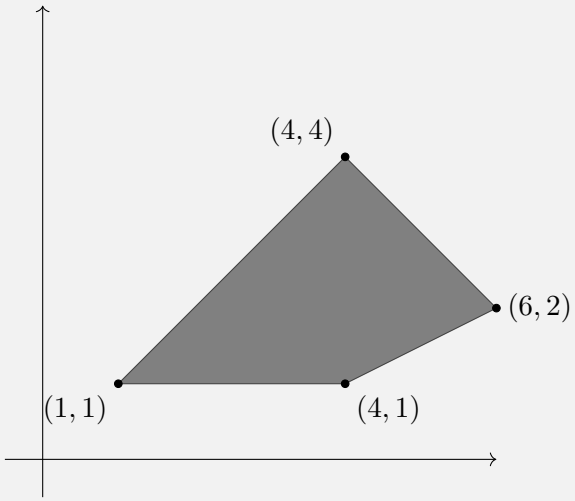
In general a region may not be expressible as a single type I or type II domain. In such a situation we can cut the region up into several pieces which are type I or type II. We can do this because of the following theorem.

**Theorem A.2.**

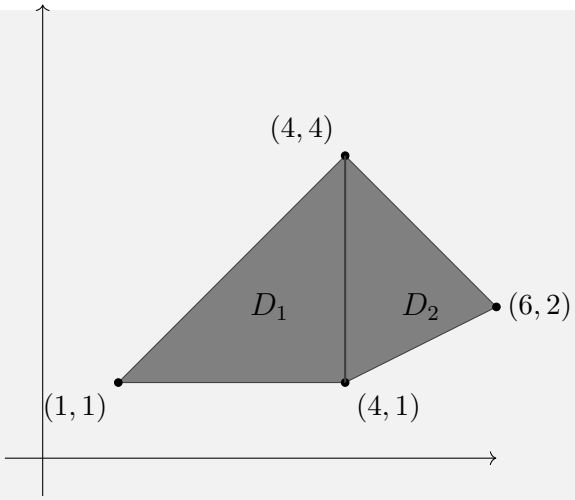
*Suppose that $D_1$ and $D_2$ are two regions in $\mathbb{R}^2$ which don't overlap. Then, writing $D = D_1 \cup D_2$,*

$$\iint_D f(x,y)\,dA = \iint_{D_1} f(x,y)\,dA + \iint_{D_2} f(x,y)\,dA.$$

**Example A.6.**
Integrate the function $x + y^2$ over the region indicated in the figure below.




We'll split this up into two regions, each of which is type I.




Our regions are

$$D_1 = \left\{ (x,y) \middle| 1 \le x \le 4,\, 1 \le y \le x \right\}$$
$$D_2 = \left\{ (x,y) \middle| 4 \le x \le 6,\, \frac{x}{2} - 1 \le y \le 8 - x \right\}.$$

Integrating over $D_1$ we have

$$\begin{aligned}
\iint_{D_1} (x+y)\, dA &= \int_1^4 \int_1^x (x+y)\, dy\, dx \\
&= \int_1^4 \left( xy + \frac{y^2}{2} \right) \Bigg|_1^x dx \\
&= \int_1^4 \left( x^2 + \frac{x^2}{2} - x - \frac{1}{2} \right) dx \\
&= \int_1^4 \left( \frac{3x^2}{2} - x - \frac{1}{2} \right) dx \\
&= \left( \frac{x^3}{2} - \frac{x^2}{2} - \frac{x}{2} \right) \Bigg|_1^4 \\
&= 32 - 8 - 2 - \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \\
&= 22.5
\end{aligned}$$

Integrating over $D_2$,

$$\begin{aligned}\iint_{D_2}(x+y)\,dA &= \int_4^6\int_{x/2-1}^{8-x}(x+y)\,dy\,dx\\ &= \int_4^6\left(xy+\frac{y^2}{2}\right)\Bigg|_{x/2-1}^{8-x}dx\\ &= \int_4^6\left[\left(8x-x^2+\frac{64-16x+x^2}{2}\right)-\left(\frac{x^2}{2}-x+\frac{x^2-4x+4}{8}\right)\right]dx\end{aligned}$$

Before integrating, let's simplify the integrand a little bit.

$$\begin{aligned}&\int_4^6\left[\left(8x-x^2+\frac{64-16x+x^2}{2}\right)-\left(\frac{x^2}{2}-x+\frac{x^2-4x+4}{8}\right)\right]dx\\ =&\frac{1}{8}\int_4^6\left[\left(64x-8x^2+256-64x+4x^2\right)-\left(4x^2-8x+x^2-4x+4\right)\right]dx\\ =&\frac{1}{8}\int_4^6\left[\left(-4x^2+256\right)-\left(5x^2-12x+4\right)\right]dx\\ =&\frac{1}{8}\int_4^6\left(-9x^2+12x+252\right)dx\end{aligned}$$

This simplified integral is much easier to integrate, but let's first notice that each term in the integrand is a multiple of 3, so we can pull the 3 out:

$$\begin{aligned}&\frac{1}{8}\int_4^6\left(-9x^2+12x+252\right)dx\\ =&\frac{3}{8}\int_4^6\left(-3x^2+4x+84\right)dx\\ =&\frac{3}{8}\left(-x^3+2x^2+84x\right)\Big|_4^6\\ =&\frac{3}{8}\left[(-216+72+504)-(-64+32+336)\right]\\ =&\frac{3}{8}(360-304)\\ =&21\end{aligned}$$

Thus

$$\begin{aligned}\iint_D (x+y)\,dA &= \iint_{D_1} (x+y)\,dA + \iint_{D_2} (x+y)\,dA \\ &= \int_1^4 \int_1^x (x+y)\,dy\,dx + \int_4^6 \int_{x/2-1}^{8-x} (x+y)\,dy\,dx \\ &= 22.5 + 21 \\ &= 43.5\end{aligned}$$

# B

# *Solutions to Exercises*

## B.1 Chapter 1

**1.1** We begin with the set of all positive multiples of 4 less than 50,

$$\{4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48\},$$

then we remove everything which is not a multiple of 6, leaving

$$\{12, 24, 36, 48\}.$$

**1.2** (a)

$$\{x \mid x = 5n \text{ for some } n \in \mathbb{N}\}$$

(b)

$$\{x \mid x = 5n \text{ for some } n \in \mathbb{Z}\}$$

**1.3** Suppose $A \subsetneq B$. That is, $A \subseteq B$ but $A \neq B$. This means $B \not\subseteq A$. Since $B$ is not a subset of $A$, it is not the case that every element of $B$ is also an element of $A$; in other words, there exists at least one element of $B$ (possibly many more, but at least one) which is not an element of $A$.

**1.4** Here's another way to think about subsets that might make $\emptyset \subseteq A$ a little easier to digest. By definition, $B \subseteq A$ if every element of $B$ is also an element of $A$. Think of this like a test: you hand me an element of $B$ and I tell you *Pass* or *Fail*, where I say *Pass* if the element is an element in $A$, and *Fail* if it's not. To see if $B \subseteq A$ or not, we'll subject *every* element of $B$ to this test. If any element of $B$ fails the test, then $B$ is not a subset of $A$. However, if no element fails the test, then $B$ is a subset of $A$.

Now, for any set $A$ let's try to apply this test to $\emptyset$. So, for every element of $\emptyset$ we apply our test, and if nothing fails, then $\emptyset \subseteq A$. There are no elements of $\emptyset$, however, so there's nothing to fail. There's no failure, so $\emptyset \subseteq A$.

## B.2 Chapter 2

**2.1** To show $A \subseteq A \cup B$, we need to show that every element of $A$ is also an element of $A \cup B$. Let $x \in A$ be any element of $A$; we need to show $x \in A \cup B$ as well. Notice, however, that $A \cup B$ consists of all elements

which are in $A$ *or* $B$. Since $x$ is in $A$, it is certainly in $A$ or $B$, and so $x \in A \cup B$. Thus $A \subseteq A \cup B$.

The argument that $B \subseteq A \cup B$ is exactly the same, but with $B$'s where $A$'s appeared above.

**2.2** By definition, $A \cap B$ contains everything that is in both $A$ and in $B$. Thus every element of $A \cap B$ is an element of $A$, and this is exactly what it means to say $A \cap B \subseteq A$. By the same token, $A \cap B \subseteq B$.

**2.3** By assumption $A \subseteq B$, and so every element of $A$ is also an element of $B$. Since $A \cap B$ consists of all the elements of both $A$ and $B$, and everything in $A$ is already an element of $B$, we see $A \cap B$ doesn't "remove" anything from $A$.

**2.4** The union of the $B_n$'s is the open interval $(0, 1)$. To see this, let's let $U$ denote the infinite union, $U = \bigcup_{n=1}^{\infty} B_n$. We want to show $U = (0, 1)$, which means we need to show $U \subseteq (0, 1)$ and $(0, 1) \subseteq U$. It is easy to see $U \subseteq (0, 1)$ since each $B_n \subseteq (0, 1)$. To see $(0, 1) \subseteq U$, let $x \in (0, 1)$ be any arbitrary element. Since $x > 0$, there exists some value of $m_1$ such that $x > \frac{1}{2^{m_1}}$ as $\frac{1}{2^n}$ decreases to 0 as $n$ increases. Notice if $x > \frac{1}{2^{m_1}}$, then $x > \frac{1}{2^M}$ for any $M > m_1$. Similarly, since $x < 1$, there exists some $m_2$ such that $x < 1 - \frac{1}{2^{m_2}}$. Note also that if $M > m_2$, then $x < 1 - \frac{1}{2^M}$.

Now let $M$ be the maximum of $m_1$ and $m_2$, $M = \max\{m_1, m_2\}$. Then $x > \frac{1}{2^M}$ and $x < 1 - \frac{1}{2^M}$; i.e., $x \in B_M$. Since $B_M \subseteq U$, this shows $x \in U$.

Thus we have established that $(0, 1) = U$.

**2.5** For notation convenience, let's write $D = E^c$ for the moment. Then $D$ is made up of all the $x \in \mathcal{U}$ such that $x \notin E$. So what is $D^c$, aka $(E^c)^c$? By definition, $D^c$ is the set of all $x \in \mathcal{U}$ such that $x \notin D$. But what does it mean if $x \notin D$? Since $D$ consists of everything *not* in $E$, if $x \notin D$ that must mean $x \in E$. That is $D^c = E$.

**2.6** The complement of the empty set, by definition, is the collection of all elements of $\mathcal{U}$ which are not elements of the empty set. But since the empty set has no elements, nothing in $\mathcal{U}$ is in the empty set, and so the complement of $\emptyset$ is the entire universe $\mathcal{U}$.

The complement of $\mathcal{U}$ is the set of all elements of $\mathcal{U}$ which are not elements of $\mathcal{U}$ – of course, there are no such elements (an element can not simultaneously be in $\mathcal{U}$ and not in $\mathcal{U}$, and so the set of all such elements is empty. I.e., $\mathcal{U}^c = \emptyset$.

**2.7** To show that two sets are equal, we need to show that each is a subset of the other. That is, we must show $E \backslash F \subseteq E \backslash (F \cap E)$ and $E \backslash (F \cap E) \subseteq E \backslash F$.

First note that if $x \in E \setminus F$, that means $x$ is in $E$ *but not* in $F$. If $x$ is not in $F$, then in particular it's not in $F \cap E$ (everything in $F \cap E$ is in $F$). Thus $x \in E \setminus (F \cap E)$, and so $E \setminus F \subseteq E \setminus (F \cap E)$.

Now suppose $x \in E \setminus (F \cap E)$. That is, $x \in E$ but $x$ is not in $F \cap E$. This means in particular that $x \notin F$: we already know $x \in E$ so if $x \in F$ as well, we would have $x \in F \cap E$. Hence $x \in E$ but $x \notin F$, which precisely means $x \in E \setminus F$. Thus $E \setminus (F \cap E) \subseteq E \setminus F$.

Together these mean that the two sets are equal.

**2.8** Let $x \in (E \cap F)^c$. This means $x \notin E \cap F$; so $x$ is not in both of $E$ and $F$ (it could be in one or the other, but it is not in their overlap). If $x$ is in neither $E$ nor $F$, then $x \in E^c$ and $x \in F^c$; both of which imply $x \in E^c \cup F^c$. If $x \in E$ but $x \notin F$, then $x \in F^c$ and so $x \in E^c \cup F^c$. Likewise, if $x \in F$ but $x \notin E$, then $x \in E^c$ so $x \in E^c \cup F^c$. This means $(E \cap F)^c = E^c \cup F^c$.

(The above is a little bit wordy, but the idea is actually simple. If you don't follow the word above, try drawing a Venn diagram and marking a point in $\mathcal{U}$ for each of the three situations above.)

Now suppose $x \in E^c \cup F^c$. That is, $x$ is in $E^c$ or $x$ is in $F^c$, or it could be in both. We again consider three cases. If $x$ is in both $E^c$ and $F^c$, that means $x$ is in neither $E$ nor in $F$, and so $x$ is *not* in $E \cup F$. In particular, since $E \cap F \subseteq E \cup F$, this means $x \notin E \cap F$ and so $x \in (E \cap F)^c$. If $x$ is in $E^c$ but not in $F^c$, then $x \notin E$ but $x \in F$. Since $x \notin E$, we must have $x \notin E \cap F$ and so $x \in (E \cap F)^c$. Similarly, if $x \notin E^c$ but $x \in F^c$ we have $x \in E$ and $x \notin F$. Since $x \notin F$, $x \notin E \cap F$, so $x \in (E \cap F)^c$.

## B.3 Chapter 3

**3.1** The preimage of $\{3, 4, 5\}$ is the empty set, $\emptyset$.

## B.4 Chapter 4

**4.1** By Exercise 2.5, we know $(E^c)^c = E$. Combining this with Proposition , we have

$$\Pr(E) = \Pr((E^c)^c) = 1 - \Pr(E^c).$$

**4.2** Let $E_1$, $E_2$, $E_3$, ... be the given non-increasing sequence. Taking the complement of each event we get a non-decreasing sequence,

$$E_1^c \subseteq E_2^c \subseteq E_3^c \subseteq \cdots .$$

By Proposition 4.9,

$$\lim_{n\to\infty} \Pr(E_n^c) = \Pr\left(\bigcup_{n=1}^{\infty} E_n^c\right).$$

By de Morgan's laws, we can rewrite each side of the above equation as

$$\lim_{n\to\infty} (1 - \Pr(E_n)) = 1 - \Pr\left(\bigcap_{n=1}^{\infty} E_n\right).$$

The limit on the left can be rewritten to give

$$1 - \lim_{n\to\infty} \Pr(E_n) = 1 - \Pr\left(\bigcap_{n=1}^{\infty} E_n\right).$$

Multiplying each side of the equation by $-1$ and then adding 1 to get the 1's to cancel gives the result,

$$\lim_{n\to\infty} \Pr(E_n) = \Pr\left(\bigcap_{n=1}^{\infty} E_n\right).$$

## B.5 Chapter 5

**5.1** (a) This is really a question of how many three letter sequences can we build which begin with `L`. If the sequence is to begin with `L`, then we have to pull `L` out first, so there's only one option for what the first letter can be in our desired sequence. For the second letter, however, we can pull out any other letter, and there are five letters left (only five and not six because we just used up the `L`). Likewise, for the third letter there are four remaining options. So the number of three letter sequences starting with `L` is $1 \cdot 5 \cdot 4 = 20$. Hence the probability of building a sequence beginning with `L` is

$$\frac{20}{120} = \frac{1}{6}.$$

(Another way to think of this is that there are six options for the first letter, and only one is `L`, so there's a 1/6 chance we'll pull out an `L`.)

(b) Using the reasoning behind part (a) again, notice that there's only one way we can pull out `LA`: first we pull out the `L` then we pull out the `A`. There's only one more letter to pick, and it can be any of the four

remaining letters. So the number of three letter sequences we can build that start with `LA` is $1 \cdot 1 \cdot 4 = 4$. Thus the probability of us constructing a sequence beginning with `LA` is

$$\frac{4}{120} = \frac{1}{30}.$$

(Another way to think about this: there are $6 \cdot 5 = 30$ different ways we can select the first two letters, but only one of them corresponds to `LA`. So regardless of what happens with the third letter, the probabilitiy we start off with `LA` is $1/30$.)

(c) Now we want to know the probability our three-letter sequence ends with `W`. Notice that we don't care what happens with the first two letters, so there are five options for the first letter (five and not six because we don't want the first letter to be `W`, since then we can't use `W` as the third letter), and similarly four options for the second letter. Hence there are $5 \cdot 4 \cdot 1 = 20$ ways to get a three-letter sequence ending in `W`, so the probability we build such a sequence is $20/120 = 1/6$.

**5.2** Let's make a table where the columns tell us the first person we pick, and the row tells us the second person.

| | |
|---|---|
| Alice | Bob |
| Alice | Cassandra |
| Alice | Danielle |
| Alice | Eric |
| Alice | Fred |
| Alice | George |
| Bob | Cassandra |
| Bob | Danielle |
| Bob | Eric |
| Bob | Fred |
| Bob | George |
| Cassandra | Danielle |
| Cassandra | Eric |
| Cassandra | Fred |
| Cassandra | George |
| Danielle | Eric |
| Danielle | Fred |
| Danielle | George |
| Eric | Fred |
| Eric | George |
| Fred | George |

Counting up the combinations we see that in fact there are 21.

**5.3** (a)

$$\binom{5}{3} = \frac{5!}{(5-3)!\,3!} = \frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} = \frac{60}{6} = 10$$

(b)

$$\binom{9}{2} = \frac{9!}{(9-2)!2!} = \frac{9 \cdot 8}{2 \cdot 1} = 36$$

Notice the second expression could be written as $\frac{9!}{7!2!}$.

(c)

$$\binom{9}{7} = \frac{9!}{(9-7)!7!}$$

Let's notice this can be written as $\frac{9!}{2!7!}$. Hence this is the same as the previous problem and equals 36.

(d)

$$\binom{37}{37} = \frac{37!}{(37-37)!37!} = \frac{37!}{0! \cdot 37!} = \frac{37!}{37!} = 1$$

Remember $0! = 1$.

(e)

$$\binom{37}{0} = \frac{37!}{37!0!} = 1.$$

**5.4** 1.

$$\binom{n}{0} = \frac{n!}{(n-0)!\,0!} = \frac{n!}{n!} = 1.$$

2.

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

3.

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

**5.5** (a) There are eight options for the first city we visit, seven options for the second, and six options for the third. Thus the number of possible itineraries is

$$\frac{8!}{(8-3)!} = 8 \cdot 7 \cdot 6 = 336$$

(b) If we're dead-set that Eindhoven must be the first city, then all we have to do is choose the other two cities. There are seven options for the second city, and six options for the third, so the number of itineraries with Eindhoven being first is

$$\frac{7!}{(7-2)!} = 7 \cdot 6 = 42$$

(c) All we care about is that we visit Helsinki, regardless of whether it's the first, second, or third stop on our trip. If it was the first city, then just as in part (b) there would be $7 \cdot 6 = 42$ options for the other two cities. Similarly, if Helsinki was the second city in our trip, then we need to choose the first city (7 options) and the third city (6 options), and again there are 42 itineriaries where Helsinki is the second city. Likewise, there would be 42 itineraries where Helsinki is the third city. In total, there are then $42 + 42 + 42 = 126$ total itineraries that include Helsinki.

An alternative way to think about this is that we need to choose the other two cities. There are $\binom{7}{2} = 21$ possibilities for the other cities *not considering the order of any of the cities*. Now to order all three cities, we multiply by $3! = 6$ to account for all possible orderings of all three cities (Helsinki and whatever the other two cities we chose were). This again gives

$$3! \cdot \binom{7}{2} = 6 \cdot 21 = 126.$$

(d) If we don't care abou the order in which we visit the cities, just the cities we visit, then there are eight cities and we choose three of them to get

$$\binom{8}{3} = 56$$

(e) From part (c) we know there are 126 *ordered* trips that involve Helsinki. To get the number of unordered trips we have to divide by 2 because the order of the other two cities *does* matter in that 126 calculation. For example, in the calculation above of 126 itineraries involving Helsinki, the itinerary *Brussels, Florence, Helsinki* is different from the itinerary *Florence, Brussels, Helsinki*. We need to divide out the ordering of these other two cities, and there are $2! = 2$ ways to order the cities, so there are

$$\frac{126}{2} = 63$$

possible unordered trips that involve Helsinki.

**5.6** (a) The event we are interested in is

$$E = \{\texttt{BBG}, \texttt{BGB}, \texttt{BGG}, \texttt{GBB}, \texttt{GBG}, \texttt{GGB}\}.$$

We compute the probability of each simple event:

$$\begin{aligned}
\Pr(\{\texttt{BBG}\}) &= \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{32} \\
\Pr(\{\texttt{BGB}\}) &= \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{24} \\
\Pr(\{\texttt{BGG}\}) &= \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{12} \\
\Pr(\{\texttt{GBB}\}) &= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{8} \\
\Pr(\{\texttt{GBG}\}) &= \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{24} \\
\Pr(\{\texttt{GGB}\}) &= \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{1}{9}
\end{aligned}$$

Now adding all of these together we have

$$\Pr(E) = \frac{3}{32} + \frac{1}{24} + \frac{1}{12} + \frac{1}{8} + \frac{1}{24} + \frac{1}{9} = \frac{143}{288} \approx 0.497$$

So just shy of half of the time, a team will contain both a boy and a girl.

(b) We could perform a computation similar to part (a), but we could also make this problem easier by considering the complement. The alternative to having at least one boy is to have all girls. The probability of the all-girl team is

$$\Pr(\{\texttt{GGG}\}) = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{2}{9}.$$

Hence the probability of a team having at least one boy is

$$1 - \frac{2}{9} = \frac{7}{9}.$$

## B.6 Chapter 6

**6.1** Simply note $E \cap F = F \cap E$. Applying the above formula to compute $\Pr(F \cap E)$ (just swapping $E$'s and $F$'s) we have

$$\Pr(F \cap E) = \Pr(F|E) \cdot \Pr(E),$$

but since $E \cap F = F \cap E$ we must have $\Pr(E \cap F) = \Pr(F \cap E)$ and so

$$\Pr(E \cap F) = \Pr(F|E) \cdot \Pr(E),$$

**6.2** (a) Let's notice that $E \cup E^c = \Omega$. Intersecting both sides of this equation with $F$ tells us

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

Now distributing the intersection we can write

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

Since $E$ and $E^c$ are disjoint, this is a disjoint union and so

$$\Pr(E \cap F) + \Pr(E^c \cap F) = \Pr(F).$$

This means

$$\Pr(E^c \cap F) = \Pr(F) - \Pr(E \cap F).$$

Dividing both sides by $\Pr(F)$ we have

$$\frac{\Pr(E^c \cap F)}{\Pr(F)} = \frac{\Pr(F) - \Pr(E \cap F)}{\Pr(F)} = 1 - \frac{\Pr(E \cap F)}{\Pr(F)} = 1 - \Pr(E|F).$$

(b) This is not true; $\Pr(E|F) \neq \Pr(E|F^c)$. For a simple counter-example, consider the sample space $\Omega = \{1, 2, 3\}$ where each simple event is equally likely. Let $E$ be the event $E = \{1, 2\}$ and $F$ the event $\{2, 3\}$. Then $\Pr(E|F) = 1/2$ while $\Pr(E|F^c) = 1$.

**6.3** Since the $F_i$ are pairwise disjoint, the $E \cap F_i$ must be disjoint as well: if $(E \cap F_i) \cap (E \cap F_j)$ was not empty, that would mean there was an element that belonged to both $F_i$ and $F_j$. This is impossible, however, since we already know that $F_i$ and $F_j$ are disjoint. Hence we must have that $E \cap F_i$ and $E \cap F_j$ are disjoint.

To show the union of the $E \cap F_i$ give $E$, simply note that

$$(E \cap F_1) \cup (E \cap F_2) \cup \cdots \cup (E \cap F_n) = E \cap (F_1 \cup F_2 \cup \cdots \cup F_n).$$

But we know $\Omega = F_1 \cup F_2 \cup \cdots \cup F_n$. Thus

$$E \cap (F_1 \cup F_2 \cup \cdots \cup F_n) = E \cap \Omega = E$$

since $E \subseteq \Omega$.

**6.4** We repeat the same Bayes' formula calcuation but now using $\Pr(D) = 0.90733$ (notice this means also $\Pr(D^c) = 0.09267$) and compute

$$\begin{aligned}\Pr(D|T) &= \frac{\Pr(T|D)\cdot\Pr(D)}{\Pr(T|D)\cdot\Pr(D)+\Pr(T|D^c)\cdot\Pr(D^c)}\\ &= \frac{0.99\cdot 0.90733}{0.99\cdot 0.090733+0.01\cdot 0.09267}\\ &= \frac{0.8983}{0.8992}\\ &= 0.99899\end{aligned}$$

So if the third test comes back positive, the chance of having the disease (assuming that the first two tests were also positive) is around 99.899%.

**6.5** Here we are told $\Pr(U\cap Y|S) = 0.4$ and $\Pr(U\cap Y|S^c) = 0.001$, and we want to compute $\Pr(S|U\cap Y)$. Applying Bayes' formula we have the following:

$$\begin{aligned}\Pr(S|U\cap Y) &= \frac{\Pr(U\cap Y|S)\cdot\Pr(S)}{\Pr(U\cap Y|S)\cdot\Pr(S)+\Pr(U\cap Y|S^c)\cdot\Pr(S^c)}\\ &= \frac{0.4\cdot 0.75}{0.4\cdot 0.75+0.001\cdot 0.25}\\ &= \frac{0.3}{0.30025}\\ &= 0.999167\end{aligned}$$

So there is a 99.9167% chance the email is spam.

**6.6** By the definition of conditional probability,

$$\Pr(E|F) = \frac{\Pr(E\cap F)}{\Pr(F)} = \frac{\Pr(\emptyset)}{\Pr(F)} = 0.$$

Since $\Pr(E) > 0$, however, $\Pr(E|F) \neq \Pr(E)$ and so the events *are not* independent.

**6.7** Consider applying Lemma 6.2. Note $\Pr(E\cap E) = \Pr(E)$, but $\Pr(E)\cdot\Pr(E) = \Pr(E)^2$. So, if $E$ is independent from itself, then Lemma 6.2 tells we must have $\Pr(E) = \Pr(E)^2$. This is only possible if $\Pr(E)$ is 0 or 1.

So, it is possible for an event to be independent of itself, but this only happens for events with probability 1 or probability 0.

**6.8**

$$\begin{aligned}\Pr(E_1 \cup E_2 \cup \cdots \cup E_n) &= 1 - \Pr([E_1 \cup E_2 \cup \cdots \cup E_n]^c) \\ &= 1 - \Pr(E_1^c \cap E_2^c \cap \cdots \cap E_n^c) \\ &= 1 - \Pr(E_1)^c \cdot \Pr(E_2)^c \cdot \ldots \cdot \Pr(E_n)^c\end{aligned}$$

## B.7 Chapter 7

**7.1** The random variable is discrete because it only takes on finitely-many values: the only outputs of the random variable are the points scored which is one of eighteen values.

## B.8 Chapter 8

**8.1** The probability $X = 1$ is the probability our first flip is a heads, which because of the way the coin is weighted we know is $2/3$. For $X$ to be equal to 2 we must have one tails and then a heads. The probability of tails is $1/3$ and the probability of heads is $2/3$, so the probability of tails then heads is $\frac{1}{3} \cdot \frac{2}{3} = \frac{2}{9}$. Similarly, if $X = 3$ we must have two tails and then a heads, and this happens with probability $\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{27}$.

In general, to get our first heads on the $n$-th flip we must first have $n-1$ tails followed by a heads, and this happens with probability $\left(\frac{1}{3}\right)^{n-1} \cdot \frac{2}{3} = \frac{2}{3^n}$. Thus the pmf is

$$p(x) = \begin{cases} 2/3^x & \text{if } x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

**8.2** We need to find the probability $X > 0$. Since the only way $X$ can be greater than 0 is for $X = 1$ or $X = 2$ which have probabilities $p(1)$ and $p(2)$, respectively, we have

$$\Pr(X > 0) = p(1) + p(2) = \frac{4}{15} + \frac{1}{5} = \frac{7}{15}.$$

**8.3** Let's first notice that $p(x)$ is equal to zero unless $x = 1$, or $x = 2$, or $x = 3$, or ... This means we can rewrite the sum as

$$\sum_{x \in \mathbb{R}} p(x) = \sum_{x=1}^{\infty} p(x).$$

Plugging in the definition of $p(x)$ for these values we have

$$\sum_{x=1}^{\infty} p(x) = \sum_{x=1}^{\infty} \frac{2}{3^x} = \sum_{x=1}^{\infty} 2 \cdot \left(\frac{1}{3}\right)^x.$$

This is *almost* something we can evaluate: if we modify this series so that it starts as $x = 0$ instead of $x = 1$ we can use the formula for a geometric series.

Recall that if $|r| < 1$, then there's a nice formula for the geometric series

$$\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}.$$

To apply this formula we need to modify our series above so that it starts at zero instead of one. To do this, let's notice that if we start the series at zero, we're adding on an extra term, namely $2 \cdot \left(\frac{1}{3}\right)^0 = 2$. This is not part of our original series, so we need to subtract it off:

$$\sum_{x=1}^{\infty} 2 \cdot \left(\frac{1}{3}\right)^x = \sum_{x=0}^{\infty} 2 \cdot \left(\frac{1}{3}\right)^x - 2.$$

Now we apply the formula for a geometric series to obtain

$$\sum_{x=0}^{\infty} 2 \cdot \left(\frac{1}{3}\right)^x - 2 = \frac{2}{1 - 1/3} - 2 = \frac{2}{2/3} - 2 = \frac{2 \cdot 3}{2} - 2 = 3 - 2 = 1.$$

**8.4**

$$p(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/18 & \text{if } 1 \leq x < 2 \\ 5/36 & \text{if } 2 \leq x < 3 \\ 7/36 & \text{if } 3 \leq x < 4 \\ 10/36 & \text{if } 4 \leq x < 5 \\ 12/36 & \text{if } 5 \leq x < 6 \\ 15/36 & \text{if } 6 \leq x < 7 \\ 17/36 & \text{if } 7 \leq x < 8 \\ 20/36 & \text{if } 8 \leq x < 9 \\ 22/36 & \text{if } 9 \leq x < 10 \\ 25/36 & \text{if } 10 \leq x < 11 \\ 27/36 & \text{if } 11 \leq x < 12 \\ 30/36 & \text{if } 12 \leq x < 14 \\ 31/36 & \text{if } 14 \leq x < 16 \\ 32/36 & \text{if } 16 \leq x < 18 \\ 33/36 & \text{if } 18 \leq x < 20 \\ 34/36 & \text{if } 20 \leq x < 22 \\ 35/36 & \text{if } 22 \leq x < 24 \\ 1 & \text{if } x \geq 24 \\ 0 & \text{otherwise} \end{cases}$$

**8.5** Recall that the probability $X = x$ (i.e., $p(x)$) is given by

$$p(x) = F(x) - \lim_{t \to x^-} F(t).$$

That is, $p(x)$ is given by the "jumps" in the cdf. Measuring these jumps in the cdf above gives the following:

$$p(x) = \begin{cases} 1/2 & \text{if } x = 0 \\ 1/10 & \text{if } x = 1 \\ 1/5 & \text{if } x = 2 \\ 1/10 & \text{if } x = 3 \\ 1/10 & \text{if } x = 3.5 \\ 0 & \text{otherwise} \end{cases}$$

**8.6** Here $X$ takes on the values 1 through 6, each with probability $1/6$. The expected value is thus

$$\mathbb{E}[X] = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 21/6 = 3/2$$

**8.7**

$$\mathbb{E}[X] = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 100 \cdot 1/6 = 115/6 \approx 19.167$$

**8.8**

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathbb{R}} x\, p(x) \\
&= -4 \cdot 7/15 + (-2) \cdot 2/15 + 0 \cdot 1/5 + 1 \cdot 1/15 + 3 \cdot 2/15 \\
&= \frac{-28 - 4 + 3 + 1 + 6}{15} \\
&= \frac{-22}{5}
\end{aligned}$$

**8.9** By Theorem 8.5, we can write

$$\begin{aligned}
\mathbb{E}[mX + b] &= \sum_{x \in \mathbb{R}} (mx + b) p(x) \\
&= \sum_{x \in \mathbb{R}} (mx\, p(x) + b\, p(x)) \\
&= m \sum_{x \in \mathbb{R}} x\, p(x) + b \sum_{x \in \mathbb{R}} p(x)
\end{aligned}$$

The first factor is simply $m$ times $\mathbb{E}[X]$, and by Corollary 8.2 the second factor is $b \cdot 1 = b$, and so

$$\mathbb{E}[mX + b] = m\mathbb{E}[X] + b.$$

**8.10**

$$\sigma_{mX+b} = \sqrt{\mathrm{Var}(mX + b)} = \sqrt{m^2 \mathrm{Var}(X)} = |m| \sqrt{\mathrm{Var}(X)} = |m| \sigma_X.$$

## B.9 Chapter 9

**9.1**

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

**9.2** Recall that the pdf for $X \sim \text{Geo}(q)$ is $p(k) = (1-q)^{k-1}q$. Note that $P(X > m)$ is then given by

$$\begin{aligned} P(X > m) &= \sum_{k=m+1}^{\infty} (1-q)^{k-1}q \\ &= \sum_{k=m}^{\infty} (1-q)^{k}q \\ &= \sum_{k=0}^{\infty} (1-q)^{m+k}q \\ &= (1-q)^m \sum_{k=0}^{\infty} (1-q)^{k}q \\ &= (1-q)^m \sum_{k=1}^{\infty} (1-q)^{k-1}q \\ &= (1-q)^m \end{aligned}$$

where the last step follows from the fact that we are summing up a pdf over all possible values and we know this sums to 1. Likewise, $P(X > n) = (1-q)^n$.

Now we simply write out the formula for conditional probability:

$$\begin{aligned} P(X > m | X > n) &= \frac{P(X > m \text{ and } X > n)}{P(X > n)} \\ &= \frac{P(X > m)}{P(X > n)} \qquad (\text{as } m > n) \\ &= \frac{(1-q)^m}{(1-q)^n} \\ &= (1-q)^{m-n} \\ &= P(X > m - n) \end{aligned}$$

**9.3**

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x\in\mathbb{R}} xp(x) \\
&= \sum_{x=0}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} xe^{-\lambda}\frac{\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} xe^{-\lambda}\frac{\lambda\cdot\lambda^{x-1}}{x(x-1)!} \\
&= e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda.
\end{aligned}$$

**9.4** We write $X^2 = X(X-1)+X$ to obtain, as before

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)+X] = \mathbb{E}[X(X+1)] + \mathbb{E}[X]$$

and we just computed $\mathbb{E}[X] = \lambda$, so now we compute the other term:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{x=0}^{\infty} x(x-1)e^{-\lambda}\frac{\lambda^x}{x!} \\
&= \lambda^2 e^{-\lambda}\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} \\
&= \lambda^2 e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \\
&= \lambda^2.
\end{aligned}$$

Thus $\mathbb{E}[X^2] = \lambda^2+\lambda$, and so

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2+\lambda-\lambda^2 = \lambda.$$

## B.10 Chapter 10

**10.1** If the pdf of $X$ is $f(x)$, then

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f(x)\, dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mathbb{E}[X] + \mathbb{E}[X]^2) f(x)\, dx \\
&= \int_{-\infty}^{\infty} x^2 f(x)\, dx - 2\mathbb{E}[X] \int_{-\infty}^{\infty} x f(x)\, dx + \mathbb{E}[X]^2 \int_{-\infty}^{\infty} f(x)\, dx \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X] \cdot \mathbb{E}[X] + \mathbb{E}[X]^2 \cdot 1 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

## B.11 Chapter 11

**11.1**

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\, dx \\
&= \int_{A}^{B} \frac{x}{B - A}\, dx \\
&= \left. \frac{x^2}{2(B - A)} \right|_{A}^{B} \\
&= \frac{B^2 - A^2}{2(B - A)} \\
&= \frac{(B + A)(B - A)}{2(B - A)} \\
&= \frac{A + B}{2}
\end{aligned}$$

**11.2**

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \int_A^B \frac{x^2}{B-A}\,dx - \left(\frac{A+B}{2}\right)^{@} \\
&= \frac{x^3}{3(B-A)}\bigg|_A^B - \frac{A^2+2AB+B^2}{4} \\
&= \frac{B^3-A^3}{3(B-A)} - \frac{A^2+2AB+B^2}{4} \\
&= \frac{(B-A)(A^2+AB+B^2)}{3(B-A)} - \frac{A^2+2AB+B^2}{4} \\
&= \frac{A^2+AB+B^2}{3} - \frac{A^2+2AB+B^2}{4} \\
&= \frac{4A^2+4AB+B^2-3A^2-6AB-3B^2}{12} \\
&= \frac{A^2-2AB+B^2}{12} \\
&= \frac{(A-B)^2}{12} \\
&= \frac{(B-A)^2}{12}
\end{aligned}$$

**11.3** We simply find the value of $\eta$ solving $F(\eta) = p$. That is,

$$\begin{aligned}
\frac{\eta - A}{B-A} &= p \\
\implies \eta &= (B-A)p + A.
\end{aligned}$$

**11.4** For the expected value we compute

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^{\infty} x\lambda e^{-\lambda x}\,dx.$$

We perform integration by parts with

$$\begin{aligned}
u &= \lambda x & dv &= e^{-\lambda x}\,dx \\
du &= \lambda\,dx & v &= \frac{-e^{-\lambda x}}{\lambda}
\end{aligned}$$

to write the integral as

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x\lambda e^{-\lambda x}\,dx \\
&= \lambda x \cdot \frac{-e^{-\lambda x}}{\lambda}\bigg|_0^\infty - \int_0^\infty \frac{-e^{-\lambda x}}{\lambda} \cdot \lambda\,dx \\
&= -xe^{-\lambda x}\bigg|_0^\infty + \int_0^\infty e^{-\lambda x}\,dx \\
&= -xe^{-\lambda x}\bigg|_0^\infty + \frac{-e^{-\lambda x}}{\lambda}\bigg|_0^\infty \\
&= \lim_{x\to\infty}\left(-xe^{-\lambda x}\right) - 0 + \lim_{x\to\infty}\frac{-e^{-\lambda x}}{\lambda} - \frac{-e^0}{\lambda}.
\end{aligned}$$

To compute the first limit we rewrite it as

$$\lim_{x\to\infty}\left(-xe^{-\lambda x}\right) = \lim_{x\to\infty}\frac{-x}{e^{-\lambda x}}$$

and then apply l'Hôpital's rule to write this as

$$\lim_{x\to\infty}\frac{-x}{e^{\lambda x}} = \lim_{x\to\infty}\frac{-1}{\lambda e^{\lambda x}} = 0.$$

The second limit in the calculation of expected value above is obviously zero, leaving

$$\mathbb{E}[X] = -\frac{-e^0}{\lambda} = \frac{1}{\lambda}.$$

To compute the variance we must calculate $\mathbb{E}[X^2]$ which proceeds similarly to the above. First, by definition,

$$\mathbb{E}[X^2] = \int_{-\infty}^\infty x^x f(x)\,dx = \int_0^\infty x^2\lambda e^{-\lambda x}\,dx.$$

We perform integration by parts with

$$\begin{aligned}
u &= \lambda x^2 & dv &= e^{-\lambda x}\,dx \\
du &= 2\lambda x\,dx & v &= \frac{-e^{-\lambda x}}{\lambda}
\end{aligned}$$

to write $\mathbb{E}[X^2]$ as

$$\lambda x^2 \cdot \frac{-e^{-\lambda x}}{\lambda}\bigg|_0^\infty - \int_0^\infty \frac{-e^{-\lambda x}}{\lambda} \cdot 2\lambda x\,dx = -x^2e^{-\lambda x}\bigg|_0^\infty + 2\int_0^\infty xe^{-\lambda x}\,dx$$

The first term is zero by an application of l'Hôpital's rule. The second term is precisely $2/\lambda$ times the integral for $\mathbb{E}[X]$ which we had calculated above as $1/\lambda$. That is,

$$\mathbb{E}[X^2] = \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2}.$$

Now we compute the variance as

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

## B.12 Chapter 12

**12.1** First we compute the marginal pmf's which are easily seen to be

$$p_X(x) = \begin{cases} 1/4 & \text{if } x \in \{1,2,3,4\} \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 1/4 & \text{if } x \in \{1,4,9,16\} \\ 0 & \text{otherwise} \end{cases}$$

The expected values are thus

$$\mathbb{E}[X] = \frac{5}{2} \quad \text{and } \mathbb{E}[Y] = \frac{15}{2}.$$

The expected value of $XY$ is

$$\mathbb{E}[XY] = \frac{1+8+27+64}{4} = 25.$$

Thus the covariance is

$$\mathrm{Cov}(X,Y) = 25 - \frac{5}{2} \cdot \frac{15}{2} = \frac{100-75}{4} = \frac{25}{4}.$$

Compared to the first part of Example **??**, this tells us that the magnitude of the covariance is related to how quickly $Y$ increases relative to $X$.

**12.2** We simply write out the definition of covariance and perform some simple algebraic manipulations to make the left-hand side look like the right-hand side.

1.

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[YX] - \mathbb{E}[Y]\mathbb{E}[X] \\ &= \mathrm{Cov}(Y,X) \end{aligned}$$

2.

$$\begin{aligned}\mathrm{Cov}(X+Y,Z) &= \mathbb{E}[(X+Y)Z] - \mathbb{E}[X+Y]\mathbb{E}[Z] \\ &= \mathbb{E}[XZ+YZ] - (\mathbb{E}[X]+\mathbb{E}[Y])\,\mathbb{E}[Z] \\ &= \mathbb{E}[XZ] + \mathbb{E}[YZ] - \mathbb{E}[X]\mathbb{E}[Z] - \mathbb{E}[Y]\mathbb{E}[Z] \\ &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] + \mathbb{E}[YZ] - -\mathbb{E}[Y]\mathbb{E}[Z] \\ &= \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)\end{aligned}$$

3.

$$\begin{aligned}\mathrm{Cov}(\lambda X,Y) &= \mathbb{E}[\lambda XY] - \mathbb{E}[\lambda X]\mathbb{E}[Y] \\ &= \lambda\mathbb{E}[XY] - \lambda\mathbb{E}[X]\mathbb{E}[Y] \\ &= \lambda\left(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\right) \\ &= \lambda\mathrm{Cov}(X,Y)\end{aligned}$$

**12.3** In the first case the correlation is 1, whereas in the second case it is $-1$.

**12.4** Simply apply Lemma 12.9 to write

$$\mathrm{Var}\left(\sum_{i=1}^{n}\lambda_i X_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\mathrm{Cov}(X_i,X_j).$$

Now note that if the $X_i$ are mutually independent, each of the terms $\mathrm{Cov}(X_i,X_j)$ is zero when $i \neq j$. Eliminating these terms and keeping the ones when $i = j$ we have

$$\sum_{i=1}^{n}\lambda_i\lambda_i\mathrm{Cov}(X_i,X_i) = \sum_{i=1}^{n}\lambda_i^2\mathrm{Var}(X_i).$$

# C

# *Solutions to Practice Problems*

## C.1 Chapter 1

**1.1** (a) $\{n^2 \,|\, n \in \mathbb{N}\}$

(b) $\{(-1)^n n^2 \,|\, n \in \mathbb{N}\}$

(c) $\{\frac{x}{y} \,|\, x, y \in \mathbb{N}, y \neq 0, x = y^3\}$

(d) $\{(x, y) \,|\, y = x^3\}$

**1.2** (a)

$$A = \{15n \,|\, n \in \mathbb{N}\}$$
$$B = \{10n \,|\, n \in \mathbb{N}\}$$
$$C = \{20n \,|\, n \in \mathbb{N}\}$$
$$D = \{30n \,|\, n \in \mathbb{N}\}$$

(b) $A$ is not a subset of $B$ because there are elements of $A$ which are not elements of $B$, such as 15.

(c) $A$ is not a subset of $C$, because there are elements of $A$ which are not elements of $C$, such as 15.

(d) $A$ is not a subset of $D$, because there are elements of $A$ which are not elements of $D$, such as 15.

(e) $B$ is not a subset of $A$, because there are elements of $B$ which are not elements of $A$, such as 20.

(f) $B$ is not a subset of $C$, because there are elements of $B$ which are not elements of $C$, such as 10.

(g) $B$ is not a subset of $D$, because there are elements of $B$ which are not elements of $D$, such as 10.

(h) $C$ is not a subset of $A$, because there are elements of $C$ which are not elements of $A$, such as 20.

(i) $C$ is a subset of $B$. Every element of $C$ is a multiple of 20 and so may be written as $20n$ for some $n \in \mathbb{N}$. However, $20 = 10 \cdot 2$ and so if we let $m = 2n$, we can also write elements of $C$ as $10m$. That is, every multiple of 20 is also a multiple of 10.

(j) $C$ is not a subset of $D$, because there are elements of $C$ which are not elements of $D$, such as 20.

(k) $D$ is a subset of $A$: every multiple of 30 is a multiple of 15 as well.

(l) $D$ is a subset of $B$: every multiple of 30 is a multiple of 10 as well.

(m) $D$ is not a subset of $C$, because there are elements of $D$ which are not elements of $C$, such as 30.

**1.3** Since the circle of radius one centered at the origin is given by the equation $x^2 + y^2 = 1$, the set of all points inside the circle is

$$A = \left\{(x, y) \,\middle|\, x^2 + y^2 \leq 1\right\}.$$

We are told that $B$ is the set

$$B = \left\{(x, y) \,\middle|\, x^2 + \frac{y^2}{4} \leq 1\right\}.$$

We want to show that $A \subseteq B$, which means that every point $(x, y) \in A$ is also in $B$. To show this is true, we need to show that the $(x, y)$-coordinates of a point in $A$ also satisfy the inequality $x^2 + \frac{y^2}{4} \leq 1$.

Notice if $x^2 + y^2 \leq 1$ (i.e., if $(x, y) \in A$), then $y^2 \leq 1 - x^2$. If we divide the left-hand by 4, that makes the left-hand side even smaller. I.e.,

$$\frac{y^2}{4} \leq y^2 \leq 1 - x^2.$$

So, if $x^2 + y^2 \leq 1$, then we know $\frac{y^2}{4} \leq 1 - x^2$. Moving the $x^2$ back to the left-hand side we have

$$x^2 + \frac{y^2}{4} \leq 1.$$

Thus if $(x, y) \in A$, then $(x, y) \in B$, and so $A \subseteq B$.

**1.4** If $A \subseteq \emptyset$, that means every element of $A$ must also be an element of $\emptyset$. Hence if $A$ has any elements, $A$ is not a subset of the empty set. This means the only subset of the empty set is empty set itself: if $A \subseteq \emptyset$, then $A = \emptyset$.

**1.5** $A$ and $B$ are not the same sets. The point $(1, 2)$ is an element of $B$, but is not an element of $A$. However, $A \subseteq B$. Every point in $A$ has the form

$$\left(x, \frac{x^2 - 1}{x - 1}\right),$$

while every point in $B$ has the form

$$(x, x + 1).$$

Given a point in $A$ written in the form above, note that as long as $x \neq 1$, the fraction in the $y$-coordinate can be written as

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

That is, with the exception of $x = 1$ (since this would result in division by zero), every point in $A$ has the same for as the points in $B$. I.e., every point of $A$ is also a point of $B$ and so $A \subseteq B$.

**1.6** No: if $A \subsetneq B$, then there must exist at least one element of $B$ which is not an element of $A$, though every element of $A$ is an element of $B$.

(If, however, it was assumed that $A \not\subseteq B$ and $B \not\subseteq A$, then it would be true that each set must contain at least one element which is not in the other set.)

## C.2 Chapter 2

**2.1** $E \cap F$ consists of elements of both $E$ and $F$. That is, integers that are multiples of both 2 and 5. Since 2 and 5 have no common divisors, the only integers which are multiples of both 2 and 5 are integers which are multiples of $2 \cdot 5 = 10$. Hence

$$E \cap F = \{10n \,\big|\, n \in \mathbb{Z}\}.$$

**2.2** Since $A \cap B$ contains everything in both $A$ and $B$, if $A \subseteq A \cap B$, this means that everything in $A$ is also in $A \cap B$ – i.e., everything in $A$ is in contained in $B$, and so $B$ is a superset of $A$: $A \subseteq B$.

**2.3** The intersection is the empty set, $\emptyset$. To see this, notice that for every $x \in \mathbb{R}$, there exists an $n$ such that $x \notin (-\infty, -n)$. For example, if $x = -7$, then $x \notin (-\infty, -8)$. Thus there are no numbers in every interval $(-\infty, -n)$, and so the intersection of all of these intervals is empty.

**2.4** This is the set of all points in the $xy$-plane whose $y$ coordinate is positive,

$$\left\{(x,y)\,\middle|\,y>0\right\}.$$

## C.3 Chapter 3

**3.1** (a) The domain is $\mathbb{R}$. The range is the set of non-negative real numbers, $\{x\in\mathbb{R}\,|\,x\geq 0\}$. The function is neither injective nor surjective.

(b) The domain and range are both $\mathbb{R}$. The function is both injective and surjective.

(c) The domain is $\mathbb{R}$, and the range is $[-1,1]$. The function is neither injective nor surjective.

(d) The domain is $\mathbb{R}$ sans the points where $\sin(x)=0$; at these points $\cot(x)$ is undefined. Thus the the domain is $\mathbb{R}\setminus\{\pi n\,|\,n\in\mathbb{N}\}$. The range is $\mathbb{R}$. The function is surjective but not injective.

(e) The domain is $[0,\infty)$ and the range is $[0,\infty)$. The function is injective, but not surjective.

## C.4 Chapter 4

**4.1** The bag contains a total of 13 coins, of these thirteen are quarters, so the probability of drawing a quarter is $^{13}/_{50}$.

**4.2** It's easier to think of the complement of what we want: the opposite of getting at least one quarter is to get no quarters. The probability of getting no quarters is the product of the probability we don't get a quarter for the first coin, and the probability we don't get a quarter for the second coin.

The probability we don't get a quarter for the first coin is the number of non-quarters over the total number of coins: $^{37}/_{50}$. For the second coind we again divide the number of non-quarters by the number of coins, but keep in mind there's one less non-quarter and one less coin since we've removed one coin. I.e., the probability the second coin is also not a quarter is $^{36}/_{49}$. Together, the probability that neither coin is a quarter is $^{37}/_{50}\cdot{}^{36}/_{49}$. Hence the probability that this event does not take place (i.e., we get at least one quarter) is

$$1-\frac{37}{50}\cdot\frac{36}{49}\approx 0.456.$$

**4.3** Notice there are $6^2 = 36$ possible ways to roll two dice:

$$\begin{array}{cccccc}
(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\
(2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\
(3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\
(4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\
(5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\
(6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6)
\end{array}$$

Of these, only four add up to nine: $(4,5)$, $(5,4)$, $(6,3)$, and $(3,6)$. So the probability of rolling exactly nine is $4/36 = 1/9$

**4.4** This problem is easier to approach by thinking of the complement: instead of computing the probability we get at least ten cents, let's compute the probability of getting less than ten cents. If we pull four coins from the coin jar, notice that if we pull a single quarter or dime, we already have at least ten cents. Likewise if we pull two nickels, we would already have at least ten cents. So the only way we're going to pull four coins and get less than ten cents is to either pull all four pennies, or one nickel and three pennies.

We could now compute probabilitiesby thinking there are twenty coins and we will choose four of them, so there are $\binom{20}{4}$ possibilities. There are $\binom{8}{4}$ ways we could choose four of the eight pennies. There are $\binom{6}{1} \cdot \binom{8}{3}$ ways we could choose one of the six nickels and three of the eight pennies. Together these give the probability of pulling out at least ten cents is

$$1 - \left( \frac{\binom{8}{4}}{\binom{20}{4}} + \frac{\binom{6}{1}\binom{8}{3}}{\binom{20}{4}} \right) = \frac{4439}{4845} \approx 0.9162.$$

We could alternatively do the calculation by keeping track of the order in which we draw coins. There are $\frac{20!}{(20-4)!} = 116280$ ways we could draw coins from the jar if order mattered.

The chance of us pulling four pennies is then

$$\frac{8}{20} \cdot \frac{7}{19} \cdot \frac{6}{18} \cdot \frac{5}{17} = \frac{1680}{116280}.$$

The chance we first pull a nickel and then three pennies is

$$\frac{6}{20} \cdot \frac{8}{19} \cdot \frac{7}{18} \cdot \frac{6}{17} = \frac{2016}{116280}.$$

Note we could also pull the nickel second, third, or fourth. These probabilities correspond to simply moving that first six in the numerator above into the second, third, or fourth position and doesn't change the actual fraction. That is, we need to multiply this probability by four; the probability we pull a nickel and three pennies (regardless of the order in which the nickel is pulled) is

$$4 \cdot \frac{2016}{116280} = \frac{8064}{116280}.$$

Together, the probability we pull less than ten cents from the coin jar is

$$\frac{1680}{116280} + \frac{8064}{116280} = \frac{9744}{116280}.$$

Thus the probability we pull at least ten cents from the coin jar is

$$1 - \frac{9744}{116280} = \frac{106536}{116280} \approx 0.9162.$$

**4.5** Notice that the probability of getting the first heads on the $n$-th flip is $1/2^n$. We are only interested in getting heads on an odd-numbered flip ($n = 1$, or $n = 3$, or $n = 5$, ...), however. Writing the $k$-th odd number as $2k - 1$, we have that the probability Alice wins on her first flip would be $\frac{1}{2^{2\cdot 1-1}} = \frac{1}{2}$; the probability Alice wins on her second flip (which would be the third flip overall) is $\frac{1}{2^{2\cdot 2-1}} = \frac{1}{8}$; and so on. The probability that Alice wins the game is thus the sum of these probabilities:

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

Notice that $\left(\frac{1}{2}\right)^{n-1} = {}^{(\frac{1}{2})^n}\!/_{1/2} = 2 \cdot \left(\frac{1}{2}\right)^n$ so the above may be written as

$$\begin{aligned}
\sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^{2k-1} &= 2\sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^{2k} \\
&= 2\sum_{k=1}^{\infty} \left(\left(\frac{1}{2}\right)^{2}\right)^{k} \\
&= 2\sum_{k=1}^{\infty} \left(\frac{1}{4}\right)^{k} \\
&= 2 \cdot \left(\sum_{k=0}^{\infty} \left(\frac{1}{4}\right)^{k} - 1\right) \\
&= 2 \cdot \left(\frac{1}{1 - {}^{1}\!/_{4}} - 1\right) \\
&= 2 \cdot \left(\frac{1}{{}^{3}\!/_{4}} - 1\right) \\
&= 2 \cdot \left(\frac{4}{3} - 1\right) \\
&= 2 \cdot {}^{1}\!/_{3} \\
&= {}^{2}\!/_{3}
\end{aligned}$$

## C.5 Chapter 5

**5.1** Notice there are 11! ways to arrange the eleven distinct tiles, but only one of them spells `DRAGONFLIES`. That is, the probability of randomly spelling `DRAGONFLIES` is ${}^{1}\!/_{11!} \approx 0.00000002505$.

**5.2** There are two ways to think about this problem. The first way is to note that there are eleven tiles, so 11! arrangements. Of these, eight spell `MATHEMATICS`: the reason for eight ways to spell `MATHEMATICS` is simply because we can swap the two `M`'s and still spell `MATHEMATICS`, and similarly swap the two `A`'s, and swap the two `T`'s. That is, we can perform $2^3 = 8$ different swaps of `M`'s, `A`'s, and `T`'s and still spell `MATHEMATICS`. Thus the probability of spelling `MATHEMATICS` is ${}^{8}\!/_{11!}$.

Another way to think of this problem is that since the tiles are not distinct, there aren't actually 11! different arrangements: this 11! is double-counting some of the arrangements – namely those where we swap the two `M`, `A`, or `T` tiles. We need to divide out by two for each possible swap. Dividing by two three times is the same as dividing by $2^3 = 8$. Thus there

are $11!/8$ possible arrangements of our letters, but only one of them spells MATHEMATICS, and so the probability is $1/(11!/8) = 8/11!$.

**5.3** (a) Notice we have six friends under consideration and are choosing three of them, so there are $\binom{6}{3} = 20$ possible groups we are considering. If Claire is going to be one of the friends going with us to the movies, then we simply need to choose two of the five remaining friends, which we can do in $\binom{5}{2} = 10$ different ways. Thus the probability Claire is one of the selected friends is

$$\frac{\binom{5}{2}}{\binom{6}{3}} = \frac{10}{20} = \frac{1}{2}$$

(b) If Erica and Fred are both going with us to the movies, we only need to choose one of the four remaining friends, which we can do in $\binom{4}{1} = 4$ different ways. Thus the probability Erica and Fred are both selected is

$$\frac{\binom{4}{1}}{\binom{6}{3}} = \frac{4}{20} = \frac{1}{5}.$$

(c) Notice that since there are three girls, there is only one way to pick three girls: choose all three of them. Likewise, there is only one way to choose all boys since there are three boys. So the probability of picking all three friends of the same gender is

$$\frac{1+1}{\binom{6}{3}} = \frac{2}{20} = \frac{1}{10}.$$

**5.4** This is another problem where it is easier to consider the complement: if you do not have at least two marbles of the same color, then all of the marbles must be different colors. That is, when you select three marbles from the urn, you must have one blue, one red, and one green. The total number of ways to get three marbles from the urn is $\binom{16}{3}$. Of these we want to consider getting one blue (of which there are $\binom{8}{1}$ ways); one green (which we can do $\binom{6}{1}$ ways); and one of the two marbles ($\binom{2}{1}$ options). That is, the probability of getting one marble of each color is

$$\frac{\binom{8}{1} \cdot \binom{6}{1} \cdot \binom{2}{1}}{\binom{16}{3}} = \frac{96}{560} = \frac{6}{35} \approx 0.1714.$$

Hence the probability of getting at least two marbles of the same color is

$$1 - \frac{6}{35} = \frac{29}{35} \approx 0.8286.$$

Alternatively, you could approach the problem as follows: supposing you get one marble of each color, you have to consider all of the permutations of colors you could see (e.g., first you get blue, then you get green, then you get red; or you could get red first, followed by green, followed by blue). Since there are three colors, there are $3! = 6$ different possible permutations of colors.

The probability of you getting a blue marble first, then a green marble, then a red marble is

$$\frac{8}{16} \cdot \frac{6}{15} \cdot \frac{2}{14} = \frac{96}{3360}.$$

However, keep in mind you have to look at all permutations of colors, so we need to multiply this value by $3! = 6$ which gives

$$\frac{576}{3360} = \frac{6}{35} \approx 0.1714,$$

and so the probability of getting at least two marbles of the same color we again see is

$$1 - \frac{6}{35} = \frac{29}{35} \approx 0.8286.$$

**5.5** We must choose two ranks from the 13 possible ranks: one rank of which we will receive three cards, and one rank of which we will receive two cards. This gives $\binom{13}{2}$ possible choices for the two ranks. However, this computation doesn't include order – i.e., it doesn't distinguish between the rank for the three-of-a-kind the rank for the two-of-a-kind. For example, if our ranks are King and Queen, it does matter whether we have three Kings and two Queens versus two Kings and three Queens. To compensate for this, we'll multiply our $\binom{13}{2}$ by two. (Equivalently, we could first choose the rank for the three-of-a-kind, then choose the rank for the two-of-a-kind: $\binom{13}{1}\binom{12}{1} = 2 \cdot \binom{13}{2}$.)

Once we've chosen the ranks, we need to choose the suits. For the rank with three cards we need to choose three of the four possible suits, which we can do in $\binom{4}{3}$ ways. For the rank with two cards we need to choose two suits, which we can do in $\binom{4}{2}$ ways. Hence the number of ways to get a full house is $2\binom{13}{2} \cdot \binom{4}{3} \cdot \binom{4}{2}$. Since there are $\binom{52}{5}$ possible five card hards, the probability of getting a full house is

$$\frac{2\binom{13}{2}\binom{4}{3}\binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2598960} = \frac{6}{4165} \approx 0.00144$$

**5.6** There are a total of twenty-six socks in the drawer and we are pulling out two of them; there are $\binom{26}{2}$ ways we can pull out two socks. Of these,

there are $\binom{12}{2}$ ways to pull out two white socks; $\binom{6}{2}$ ways to pull out two black socks; $\binom{4}{2}$ ways to pull out two brown socks; and $\binom{4}{2}$ ways to pull out two blue socks. That is, the probability we pull out two socks of the same color is

$$\frac{\binom{12}{2}+\binom{6}{2}+\binom{4}{2}+\binom{4}{2}}{\binom{26}{2}} = \frac{66+15+6+6}{325} = \frac{93}{325} \approx 0.2862$$

and so there's about a 28.62% chance we would pull out two socks of the same color.

**5.7** There are twenty students and we want to choose three of them: there are $\binom{20}{3}$ ways to do this. If exactly one of our students has an Android phone and not all students have the same phone, then it must be that the other two students have iPhones. We want to select one of the eight Android users, which we can do in $\binom{8}{1}$ ways, and two of the twelve iPhone users, which we can do in $\binom{12}{2}$ ways. Thus the probability fo selecting one Android user and two iPhone users is

$$\frac{\binom{8}{1}\cdot\binom{12}{2}}{\binom{20}{3}} = \frac{44}{95} \approx 0.4632.$$

## C.6 Chapter 6

**6.1** The probability we pull out two blue socks, given that we've pulled out two socks of the same color is

$$\begin{aligned} P(\text{Two blue}|\text{Same Color}) &= \frac{P(\text{Two blue} \cap \text{Same color})}{P(\text{Same color})} \\ &= \frac{P(\text{Two blue})}{P(\text{Same color})} \\ &= \frac{\binom{4}{2}/\binom{26}{2}}{93/325} \\ &= \frac{6/325}{93/325} \\ &= \frac{6}{93} \\ &= \frac{2}{31} \approx 0.0645 \end{aligned}$$

and so there is about a 6.45% chance we pull out two blue socks, given that we pull out two socks of the same color.

**6.2** Note that

$$\begin{aligned} & P(E \cup F) = P(E) + P(F) - P(E \cap F) \\ \implies & 1/2 = 2/5 + 3/10 - P(E \cap F) \\ \implies & P(E \cap F) = 2/5 + 3/10 - 1/2 = 1/5. \end{aligned}$$

We now compute

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{1/5}{3/10} = 2/3.$$

**6.3** Notice that $P(E) = 50/100 = 1/2$ and $P(F) = 20/100 = 1/5$. The probability both events take place would be that we pull out a marble which is both a multiple of two and a multiple of five – that is, the marble would need to be a multiple of 10. Hence $P(E \cap F) = 10/100 = 1/10$. Since $P(E)P(F) = 1/10$ as well, the events are independent.

**6.4** Let $S$ be the event the student plays soccer, $F$ the event they play football, and $B$ the event they play basketball. We are told in the problem that $P(S) = 0.3$, $P(F) = 0.1$, and $P(B) = 0.25$, as well as the conditional probabilities $P(B|F \cap S) = 0.05$ and $P(F|S) = 0.1$. We now compute

$$\begin{aligned} P(B \cap F \cap S) &= P(B \cap (F \cap S)) \\ &= P(B|F \cap S) \cdot P(F \cap S) \\ &= P(B|F \cap S) \cdot P(F|S) \cdot P(S) \\ &= 0.05 \cdot 0.1 \cdot 0.3 \\ &= 0.0015 \end{aligned}$$

And so there is a 0.15% chance a randomly selected student plays all three sports.

**6.5** Let $E$ be the event a randomly selected person from the UK is English; $S$ the event they're Scottish; $I$ the event they're Irish; $W$ the event they're Welsh; and $R$ the event they have red hair. We are trying to determine if $P(E|R)$, $P(S|R)$, $P(I|R)$, or $P(W|R)$ is most likely. We are told in the problem that $P(E) = 0.6$, $P(S) = 0.2$, $P(I) = 0.15$, $P(W) = 0.05$, as well as $P(R|E) = 0.15$, $P(R|S) = 0.75$, $P(R|I) = 0.65$, and $P(R|W) = 0.3$. Notice that we can compute $P(E|R)$ as

$$P(E|R) = \frac{P(E \cap R)}{P(R)} = \frac{P(R \cap E)}{P(R)} = \frac{P(R|E)P(E)}{P(R)}$$

and similarly for $P(S|R)$, $P(I|R)$, and $P(W|R)$. (This is simply Bayes' formula.)

We can compute $P(R)$ by the law of total probability:

$$\begin{aligned} P(R) &= P(R|E)P(E) + P(R|S)P(S) + P(R|I)P(I) + P(R|W)P(W) \\ &= 0.15 \cdot 0.6 + 0.75 \cdot 0.2 + 0.65 \cdot 0.15 + 0.3 \cdot 0.05 \\ &= 0.3525 \end{aligned}$$

Now by Bayes' formula we have

$$\begin{aligned} P(E|R) &= \frac{P(R|E)P(E)}{P(R)} = \frac{0.15 \cdot 0.6}{0.3525} = \frac{0.09}{0.3525} \approx 0.2553 \\ P(S|R) &= \frac{P(R|S)P(S)}{P(R)} = \frac{0.75 \cdot 0.2}{0.3525} = \frac{0.15}{0.3525} \approx 0.4255 \\ P(I|R) &= \frac{P(R|I)P(I)}{P(R)} = \frac{0.65 \cdot 0.15}{0.3525} = \frac{0.0975}{0.3525} \approx 0.2766 \\ P(W|R) &= \frac{P(R|W)P(W)}{P(R)} = \frac{0.3 \cdot 0.05}{0.3525} = \frac{0.015}{0.3525} \approx 0.0426 \end{aligned}$$

So, a randomly selected redhead has about an 25.53% chance of being English, a 42.55% chance of being Scottish, a 27.66% chance of being Irish, and a 4.26% chance of being Welsh. So the redhead is most likely to be Scottish, and second most likely to be Irish.

**6.6** (a) Let $C$ be the event a math major is enrolled in the course, $F$ the event they are a freshmen, $So$ the event they're a sophomore, $J$ the event they're a junior, and $Se$ the event they're a senior. By the law of total probability we have

$$\begin{aligned} P(C) &= P(C|F)P(F) + P(C|So)P(So) + P(C|J)P(J) + P(C|Se)P(Se) \\ &= 0 \cdot 0.2 + {}^1\!/_3 \cdot {}^3\!/_{10} + {}^1\!/_2 \cdot {}^1\!/_4 + {}^1\!/_4 \cdot {}^1\!/_4 \\ &= {}^1\!/_{10} + {}^1\!/_8 + {}^1\!/_{16} \\ &= {}^{16}\!/_{160} + {}^{20}\!/_{160} + {}^{10}\!/_{160} \\ &= {}^{46}\!/_{160} = {}^{23}\!/_{80}. \end{aligned}$$

(b) By Bayes' formula we simply compute

$$\begin{aligned} P(J|C) &= \frac{P(C|J)P(J)}{P(C)} \\ &= \frac{1/2 \cdot 1/4}{23/80} \\ &= \frac{1/8}{23/80} \\ &= \frac{80/8}{23} \\ &= 10/23 \end{aligned}$$

**6.7** Let $F$ be the event that a random graduate works in finance, and $M$ the event they were a math major, $C$ the event they were a computer science major, $P$ the event they were a physics major, and $O$ the event they majored in something else. In the problem we are told $P(F|M) = 0.15$, $P(F|C) = 0.05$, $P(F|P) = 0.1$, and $P(F|O) = 0.03$, as well as $P(M) = 0.05$, $P(C) = 0.1$, $P(P) = 0.03$, and so we must have $P(O) = 0.82$.

To compute the probability someone working in finance was a mathematics major, we want $P(M|F)$ which we can compute with Bayes formula:

$$\begin{aligned} P(M|F) =& \frac{P(M \cap F)}{P(F)} \\ =& \frac{P(F \cap M)}{P(F)} \\ =& \frac{P(F|M)P(M)}{P(F|M)P(M) + P(F|C)P(C) + P(F|P)P(P) + P(F|O)P(O)} \\ =& \frac{0.15 \cdot 0.05}{0.15 \cdot 0.05 + 0.05 \cdot 0.1 + \cdot 0.1 \cdot 0.03 + 0.03 \cdot 0.82} \\ =& \frac{0.0075}{0.0401} \\ \approx& 0.187 \end{aligned}$$

So there is about an 18.7% chance the financier was a mathematics major.

**6.8** Let $T$ be the event that we buy a training flat, and let $S$ be the event the shoe was made in South Korea, $A$ the event it was made in Australia, and $V$ the event it was made in Venezuela. We want to find the probability the shoe was produced in South Korea given that we buy a training flat:

$P(S|T)$. By Bayes' formula we can write this as

$$\begin{aligned} P(S|T) &= \frac{P(T|S)P(S)}{P(T)} \\ &= \frac{P(T|S)P(S)}{P(T|S)P(S) + P(T|A)P(A) + P(T|V)P(V)} \\ &= \frac{0.5 \cdot 0.6}{0.5 \cdot 0.6 + 0.75 \cdot 0.2 + 0.6 \cdot 0.2} \\ &= \frac{0.3}{0.57} \\ &= \frac{30}{57} = \frac{10}{19} \\ &\approx 0.5263 \end{aligned}$$

## C.7 Chapter 7

**7.1** The random variable is continuous because it can, in principle, take on any value in the interval $[0, \infty)$. Even if there were an upper bound on how far the ball could roll (e.g., if the ball could not roll more than ten feet), the random variable would still be continuous (in the case of the ball rolling at most ten feet, the range of the random variable would be $[0, 10]$).

**7.2** The random variable is discrete because the range of the random variable is infinite, but the values in this range have a well-defined first, second, third, and so on. The first value is 0, the second value is 1, the third value is 2, etc.

**7.3** The random variable is continuous because the range, the set of all values you could potentially realize as an output of the random variable, is the interval $[0, 100]$.

## C.8 Chapter 8

**8.1** To solve this problem we want to make use of part (3) of Theorem 8.3 which says that for $F$ to be a cdf, we must have $\lim_{x \to \infty} F(x) = 1$. As $x$ goes to infinity, the value of $F(x)$ is made up of more and more and more terms, and in the limit we have

$$\lim_{x \to \infty} F(x) = \sum_{j=0}^{\infty} \frac{k}{3^j}.$$

This is a geometric series, however, and so this is equal to

$$\sum_{j=0}^{\infty} \frac{k}{3^j} = \frac{k}{1 - 1/3} = \frac{3k}{2}.$$

For this to equal one, we require $k = 2/3$.

**8.2 Solution**

$$\mathbb{E}[X - \mu] = \mathbb{E}[X] - \mu = \mu - \mu = 0.$$

**8.3**

$$\begin{aligned}
\mathbb{E}[f(X)] &= f(1) \cdot 1/10 + f(2) \cdot 1/5 + f(3) \cdot 2/5 + f(4) \cdot 3/10 \\
&= 0 + (-2) \cdot 1/5 + (-2) \cdot 2/5 + 0 \\
&= (-2) \cdot 3/5 \\
&= -6/5
\end{aligned}$$

**8.4** From the CDF we can recover the PDF as

$$p(x) = F(x) - \lim_{y \to x^-} F(y).$$

I.e., the pdf is given by the "jumps" in the CDF. Here this gives us

$$p(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 1/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 3 \\ 1/4 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

Now we compute the expected value:

$$\begin{aligned}
\mathbb{E}[X] &= 0 \cdot 1/4 + 1 \cdot 1/8 + 3 \cdot 3/8 + 5 \cdot 1/4 \\
&= 0 + 1/8 + 9/8 + 5/4 \\
&= 20/8 \\
&= 5/2
\end{aligned}$$

**8.5** From the cdf we can recover the pmf as

$$p(x) = F(x) - \lim_{y \to x^-} F(y).$$

I.e., the pmf is given by the "jumps" in the CDF. Here this gives us

$$p(x) = \begin{cases} 1/4 & \text{if } x = 0 \\ 1/8 & \text{if } x = 1 \\ 3/8 & \text{if } x = 3 \\ 1/4 & \text{if } x = 5 \\ 0 & \text{otherwise} \end{cases}$$

**8.6** We create a table of all sums of rolls of two dice. In the table below the label of the rows tell us the value of one of the dice, and the label of the columns tell us the value of the other die.

| | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **1** | 2 | 3 | 4 | 5 | 6 | 7 |
| **2** | 3 | 4 | 5 | 6 | 7 | 8 |
| **3** | 4 | 5 | 6 | 7 | 8 | 9 |
| **4** | 5 | 6 | 7 | 8 | 9 | 10 |
| **5** | 6 | 7 | 8 | 9 | 10 | 11 |
| **6** | 7 | 8 | 9 | 10 | 11 | 12 |

We can now determine the pmf of the random variable by counting the number of times each sum occurs in this table of 36 possible sums.

$$p(x) = \begin{cases} 1/36 & \text{if } x = 2 \\ 2/36 & \text{if } x = 3 \\ 3/36 & \text{if } x = 4 \\ 4/36 & \text{if } x = 5 \\ 5/36 & \text{if } x = 6 \\ 6/36 & \text{if } x = 7 \\ 5/36 & \text{if } x = 8 \\ 4/36 & \text{if } x = 9 \\ 3/36 & \text{if } x = 10 \\ 2/36 & \text{if } x = 11 \\ 1/36 & \text{if } x = 12 \\ 0 & \text{otherwise} \end{cases}$$

Now we compute the expected value:

$$\begin{aligned} \mathbb{E}[X] &= 2 \cdot 1/36 + 3 \cdot 2/36 + 4 \cdot 3/36 + \cdots + 10 \cdot 3/36 + 11 \cdot 2/36 + 12 \cdot 1/36 \\ &= 7 \end{aligned}$$

We also compute the expected value of the square,

$$\begin{aligned}\mathbb{E}[X^2] &= 2^2 \cdot {}^1\!/_{36} + 3^2 \cdot {}^2\!/_{36} + 4^2 \cdot {}^3\!/_{36} + \cdots + 10^2 \cdot {}^3\!/_{36} + 11^2 \cdot {}^2\!/_{36} + 12^2 \cdot {}^1\!/_{36} \\ &= {}^{329}\!/_{6}\end{aligned}$$

And finally we compute the variance:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{329}{6} - 49 = \frac{287}{6} \approx 47.8333.$$

**8.7** (a) To make the pmf easier to compute, let's suppose we can distinguish different sides of the same color. For example, suppose the sides on the first die are $R_1$, $R_2$, $R_3$, $B_1$, $B_2$, and $G$; and the sides on the second die are $R_1$, $R_2$, $B_1$, $B_2$, $G_1$, $G_2$. Now let's consider a table listing all 36 possible rolls of these dice with distinguishable colors. We'll list these as ordered pairs where the first entry tells us the outcome of the first die, and the second entry tells us the outcome of the second die.

$$\begin{gathered}(R_1, R_1),\ (R_1, R_2),\ (R_1, B_1),\ (R_1, B_2),\ (R_1, G_1),\ (R_1, G_2) \\ (R_2, R_1),\ (R_2, R_2),\ (R_2, B_1),\ (R_2, B_2),\ (R_2, G_1),\ (R_2, G_2) \\ (R_3, R_1),\ (R_3, R_2),\ (R_3, B_1),\ (R_3, B_2),\ (R_3, G_1),\ (R_3, G_2) \\ (B_1, R_1),\ (B_1, R_2),\ (B_1, B_1),\ (B_1, B_2),\ (B_1, G_1),\ (B_1, G_2) \\ (B_2, R_1),\ (B_2, R_2),\ (B_2, B_1),\ (B_2, B_2),\ (B_2, G_1),\ (B_2, G_2) \\ (G, R_1),\ (G, R_2),\ (G, B_1),\ (G, B_2),\ (G, G_1),\ (G, G_2)\end{gathered}$$

Now notice that in this table, where we've distinguished the sides, each of the thirty-six outcomes is equally likely. We could now replace each entry in this table by the corresponding score:

$$\begin{gathered}10,\ 10,\ 4,\ 4,\ 8,\ 8 \\ 10,\ 10,\ 4,\ 4,\ 8,\ 8 \\ 10,\ 10,\ 4,\ 4,\ 8,\ 8 \\ 4,\ 4,\ 10,\ 10,\ 3,\ 3 \\ 4,\ 4,\ 10,\ 10,\ 3,\ 3 \\ 8,\ 8,\ 3,\ 3,\ 10,\ 10\end{gathered}$$

Notice in this table 10 occurs twelve times, 8 occurs eight times, 4 occurs ten times, and 3 occurs six times. Thus, the pmf is

$$p(x) = \begin{cases} 6/36 & \text{if } x = 3 \\ 10/36 & \text{if } x = 4 \\ 8/36 & \text{if } x = 8 \\ 12/36 & \text{if } x = 10 \\ 0 & \text{otherwise.} \end{cases}$$

(b) Using the pmf from part (a) we have

$$\mathbb{E}[X] = 3 \cdot 6/36 + 4 \cdot 10/36 + 8 \cdot 8/36 + 10 \cdot 12/36 = \frac{242}{36} \approx 6.722.$$

**8.8** If the insurance company charges each customer a premium of $P$ dollars, then the company's profit for that customer is $P$ minus any money the company spends covering the customer's claims. Since there is a \$500 deductible, the customer pays the first \$500 of a claim and the company pays the remainder. E.g., if the customer is in a \$5000 accident, then the company pays \$4500 for the accident. This means the company's profit for such a customer is $P - 4500$. Note that for the trivial accidents the company simply gains a profit of $P$ from the customer since nothing is paid in the accident. Thus the expected profit with premium $P$ is

$$\begin{aligned} & P \cdot 0.8 + (P - 500) \cdot 0.1 + (P - 4500) \cdot 0.08 + (P - 9500) \cdot 0.02 \\ =& P - 50 - 360 - 190 \\ =& P - 500 \end{aligned}$$

If this is to be \$100, then the company needs to charge a premium of \$600.

**8.9** We simply plug into the formula for expected value to obtain

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x p(x) = \sum_{n=1}^{\infty} 2^n \cdot 1/2^n = \sum_{n=1}^{\infty} 1 = \infty.$$

(Note that in some books the expected value is required to be finite when it exists, and by that definition this random variable's expectation does not exist.)

**8.10** The expected revenue from a random vehicle under the toll is

$$\$1 \cdot 0.6 + \$2.5 \cdot 0.4 = \$0.6 + \$1 = \$1.6.$$

Thus the expected revenue from 25 random cars is

$$25 \cdot \$1.6 = \$40.$$

## C.9 Chapter 9

**9.1** Notice first that this is a binomial random variable with $n = 5$ trials (the questions) and probability of success (guessing the right answer) $q = 1/3$: $X \sim \text{Binomial}(5, 1/3)$. The pdf of this random variable is thus

$$p(x) = \begin{cases} \binom{5}{x} (1/3)^x (2/3)^{5-x} & \text{if } x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{otherwise} \end{cases}$$

We then compute

$$\begin{aligned} P(X \geq 4) &= P(X = 4) + P(X = 5) \\ &= p(4) + p(5) \\ &= \binom{5}{4} (1/3)^4 (2/3)^1 + \binom{5}{5} (1/3)^5 (2/3)^0 \\ &= 5 \cdot 1/81 \cdot 2/3 + 1 \cdot 1/243 \cdot 1 \\ &= \frac{10}{243} + \frac{1}{243} \\ &= \frac{11}{243} \end{aligned}$$

**9.2** Notice the number of acceptable piston heads in the sample is given by a hypergeometric random variable $X$ where the population size is $N = 50$, the sample size is $n = 8$, and the number of successes (acceptable piston heads) in the population is $k = 47$. The probability we select exactly six acceptable piston heads is given by plugging in $x = 6$ into the pdf for the hypergeometric random variable:

$$\begin{aligned} P(X = 6) &= p(6) \\ &= \frac{\binom{47}{6}\binom{3}{2}}{\binom{50}{8}} \\ &= \frac{3}{50} = 0.06 \end{aligned}$$

**9.3** Let $X$ be the number of corrupted bits, so $X \sim \text{Binomial}(7, 1/10)$. The original message can be reconstructed if $X = 0$ or $X = 1$, and this happens with probability

$$\begin{aligned} \Pr(X = 0) + \Pr(X = 1) &= \binom{7}{0} (1/10)^0 (9/10)^7 + \binom{7}{0} (1/10)^0 (9/10)^7 \\ &= \frac{531441}{625000} \\ &\approx 0.8503 \end{aligned}$$

So there's about an 85% chance the original message can be reconstructed.

## C.10 Chapter 10

**10.1** By differentiating over each interval we have

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{8} & \text{if } 0 < x < 2 \\ 0 & \text{if } 2 < x < 4 \\ \frac{1}{4} & \text{if } 4 < x < 7 \\ 0 & \text{if } x > 7 \\ 0 & \text{otherwise} \end{cases}$$

**10.2** (a) It's clear that $f(x) \geq 0$ for all $x$, so we only need to verify that it integrates to 1:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \int_{10}^{\infty} \frac{10}{x^2}\,dx \\ &= \lim_{b\to\infty} \int_{10}^{b} 10x^{-2}\,dx \\ &= \lim_{b\to\infty} \left.\frac{-10}{x}\right|_{10}^{b} \\ &= \lim_{b\to\infty} \left(\frac{-10}{b} - \frac{-10}{10}\right) \\ &= \lim_{b\to\infty} \left(\frac{-10}{b} + 1\right) \\ &= 1 \end{aligned}$$

(b) It's clear that $F(x) = 0$ for $x < 10$. For $x > 10$ we compute

$$\begin{aligned} F(x) &= \int_{10}^{x} \frac{10}{t^2}\,dt \\ &= \left.\frac{-10}{t}\right|_{10}^{x} \\ &= \frac{-10}{x} - \frac{-10}{10} \\ &= 1 - \frac{10}{x} \end{aligned}$$

and thus

$$F(x) = \begin{cases} 0 & \text{if } x < 10 \\ 1 - \frac{10}{x} & \text{if } x \geq 10 \end{cases}$$

**10.3** (a) Since the pdf must integrate to 1 we have

$$\begin{aligned} &\int_0^2 k(4x - 2x^2)\,dx = 1 \\ \implies &k\left(2x^2 - \frac{2x^3}{3}\right)\Big|_0^2 = 1 \\ \implies &k(8 - {}^{16}\!/_3) = 1 \\ \implies &\frac{8k}{3} = 1 \\ \implies &k = \frac{3}{8}. \end{aligned}$$

To see that $f(x) \geq 0$ for this choice of $k$, notice that $f(0) = f(2) = 0$, and for $x$ between 0 and 2 we have $f'(x) = 6(1 - x)$ which has a root at $x = 1$. To the left of this (on the interval $(0, 1)$) we thus see that $f$ is increasing; to the right (on $(1, 2)$) the function is decreasing, but it doesn't intersect the $x$-axis until $x = 2$. Hence the function is non-negative for all $x$.

(b)

$$\begin{aligned} P({}^1\!/_2 < X < {}^3\!/_2) &= \int_{1/2}^{3/2} \frac{3}{8}(4x - 2x^2)\,dx \\ &= \frac{3}{8}\left(2x^2 - \frac{2x^3}{3}\right)\Big|_{1/2}^{3/2} \\ &= \frac{3}{8}\left[\left(\frac{9}{2} - \frac{9}{4}\right) - \left(\frac{1}{2} - \frac{1}{12}\right)\right] \\ &= \frac{3}{8}\left[\frac{9}{4} - \frac{5}{12}\right] \\ &= \frac{3}{8} \cdot \frac{11}{6} \\ &= \frac{11}{16} \\ &= 0.6875 \end{aligned}$$

(c)

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^2 x\frac{3}{8}\left(4x - 2x^2\right)\,dx \\
&= \int_0^2 \frac{3}{8}\left(4x^2 - 2x^3\right)\,dx \\
&= \frac{3}{8}\left(\frac{4x^3}{3} - \frac{x^4}{2}\right)\Big|_0^2 \\
&= \frac{3}{8}\left(\frac{32}{3} - 8\right) \\
&= \frac{3}{8}\cdot\frac{8}{3} \\
&= 1
\end{aligned}$$

**10.4** By differentiating over each interval we have

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{8} & \text{if } 0 < x < 2 \\ 0 & \text{if } 2 < x < 4 \\ \frac{1}{4} & \text{if } 4 < x < 7 \\ 0 & \text{if } x > 7 \\ 0 & \text{otherwise} \end{cases}$$

**10.5** By the definition of conditional probability, this is equal to

$$\frac{P(X \leq 2/3 \cap X \geq 1/2)}{P(X \geq 1/2)} = \frac{P(1/2 \leq X \leq 2/3)}{P(X \geq 1/2)}$$

Now we just compute each of these probabilities by integrating the pdf:

$$\begin{aligned}
P(1/2 \leq X \leq 2/3) &= \int_{1/2}^{2/3} 4x^3\,dx \\
&= x^4\Big|_{1/2}^{2/3} \\
&= \frac{16}{81} - 1/16
\end{aligned}$$

$$\begin{aligned} P(X \geq 1/2) &= \int_{1/2}^{1} 4x^3 \, dx \\ &= x^4 \Big|_{1/2}^{1} \\ &= 1 - 1/16 \\ &= \frac{15}{16} \end{aligned}$$

Together these give that the desired probability is $\frac{35}{243} \approx 0.144$.

**10.6** (a) The median of $X$ is the 50-th percentile, so we need to find the value of $T$ such that $\int_{-\infty}^{T} f(x) \, dx = 1/2$.

Since the pdf is only supported in $[0, 1]$, it's clear this value of $T$ must occur in $[0, 1]$ and so we have

$$\begin{aligned} & \int_{-\infty}^{T} f(x) \, dx = \frac{1}{2} \\ \implies & \int_{0}^{T} 2(1 - x) \, dx = \frac{1}{2} \\ \implies & \left(2x - x^2\right) \Big|_{0}^{T} = \frac{1}{2} \\ \implies & 2T - T^2 = \frac{1}{2} \end{aligned}$$

That is, our $T$ is a solution to the quadratic $T^2 - 2T + 1/2 = 0$. Using the quadratic formula, we have

$$T = \frac{2 \pm \sqrt{4 - 4 \cdot 1 \cdot 1/2}}{2} = \frac{2 \pm \sqrt{2}}{2}.$$

Of these two solutions, however, only $\frac{2-\sqrt{2}}{2}$ is in the interval $[0, 1]$, and so our median is $T = \frac{2-\sqrt{2}}{2}$.

(b) We compute $\text{Var}(X)$ using the formula $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, so first we need to compute these expected values.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^1 2x(1-x)\,dx \\
&= \int_0^1 (2x - 2x^2)\,dx \\
&= \left(\frac{2x^2}{2} - \frac{2x^3}{3}\right)\Big|_0^1 \\
&= (1 - {}^2\!/_3) - (0-0) \\
&= {}^1\!/_3
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^1 2x^2(1-x)\,dx \\
&= \int_0^1 (2x^2 - 2x^3)\,dx \\
&= \left(\frac{2x^3}{3} - \frac{2x^4}{4}\right)\Big|_0^1 \\
&= ({}^2\!/_3 - {}^1\!/_2) - (0-0) \\
&= {}^4\!/_6 - {}^3\!/_6 \\
&= {}^1\!/_6
\end{aligned}$$

Now we compute

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = {}^1\!/_6 - {}^1\!/_9 = \frac{3-2}{18} = {}^1\!/_{18}$$

# C.11 Chapter 11

**11.1** (a) Notice if $|X| > 1/2$, that means that either $X > 1/2$ or $X < -1/2$. We thus compute

$$\begin{aligned} P(|X| > 1/2) &= P(X < -1/2) + P(X > 1/2) \\ &= \int_{-1}^{-1/2} \frac{dx}{2} + \int_{1/2}^{1} \frac{dx}{2} \\ &= \frac{x}{2}\Big|_{-1}^{-1/2} + \frac{x}{2}\Big|_{1/2}^{1} \\ &= -1/4 - (-1/2) + 1/2 - 1/4 \\ &= 1 - 1/2 \\ &= 1/2 \end{aligned}$$

(b) Notice that for $0 \leq x \leq 1$, $|X| < x$ means $-x < X < x$. Hence, assuming $0 \leq x \leq 1$, we have

$$P(|X| < x) = \int_{-x}^{x} \frac{dt}{2} = \frac{t}{2}\Big|_{-x}^{x} = \frac{x}{2} - \frac{-x}{2} = x.$$

Clearly for $x < 0$ we have $P(|X| < x) = 0$ and for $x > 1$ we have $P(|X| < x) = 1$. Thus the cdf is

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

(Notice this means $|X| \sim \text{Uni}([0, 1])$.)

**11.2** If $X$ is the random variable representing the number of accidents, then we are told in the problem $X \sim N(45, 10)$. We want to compute $P(X > 60)$, and to do this we will consider the complementary event $X \leq 60$ which we can find by transforming $X$ into the standard normal random variable $Z$ and using the table of $\Phi$ values on the first page of the exam.

$$\begin{aligned} P(X > 60) &= 1 - P(X \leq 60) \\ &= 1 - P\left(\frac{X-45}{10} \leq \frac{60-45}{10}\right) \\ &= 1 - P(Z \leq {}^{15}/_{10}) \\ &= 1 - \Phi(1.5) \\ &= 1 - 0.9332 \\ &= 0.0668 \end{aligned}$$

**11.3** We are trying to find the 70-th percentile of the random variable $X \sim N(8, 0.5)$. That is, we want to find the value of $x$ such that $P(X \leq x) = 0.7$. To compute this we will transform $X$ into the standard normal $Z \sim N(0, 1)$, whose 70-th percentile we can look up in the table on the first page of the exam.

The transformation to the standard normal is given by $Z = \frac{X-8}{0.5}$. Once transformed to the standard normal we wish to find the value of $\eta$ such that $P(Z \leq \eta) = 0.7$. We denote the cdf of the standard normal by $\Phi$, so this is the same as finding the value of $\eta$ such that $\Phi(\eta) = 0.7$. From the table on the first page of the exam we see that $\Phi(0.5244) = 0.7$, and hence the corresponding $\eta$ for the standard normal is 0.5244. We then transform this to the corresponding value for the original random variable $X$:

$$\begin{aligned} &\frac{x-8}{0.5} = 0.5244 \\ \implies &x - 8 = 0.5 \cdot 0.5244 = 0.2622 \\ \implies &x = 8.2622 \end{aligned}$$

So the 70-th percentile of birth weights in the US is 8.2622 pounds.

# C.12 Chapter 12

**12.1** (a)

$$p(0,-2) + p(0,-1) + p(1,-2) + P(1,-1) + p(2,-2) + p(1,-1) = 0.29$$

(b) For each $x$ value we simply add up all possible $y$ values.

$$p_X(x) = \begin{cases} 0.2 & \text{if } x = 0 \\ 0.2 & \text{if } x = 1 \\ 0.2 & \text{if } x = 2 \\ 0.2 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

(c) For each $y$ value we simply add up all possible $x$ values.

$$p_Y(y) = \begin{cases} 0.27 & \text{if } y = -2 \\ 0.17 & \text{if } y = -1 \\ 0.22 & \text{if } y = 1 \\ 0.14 & \text{if } y = 2 \end{cases}$$

(d) For each $(x, y)$ pair in the table we multiply the $x$ and $y$ values and the corresponding probability and add them all together:

$$\begin{aligned} \mathbb{E}[XY] =& 0 \cdot -2 \cdot 0.01 + 0 \cdot -1 \cdot 0 + 0 \cdot 1 \cdot 0.05 + 0 \cdot 2 \cdot 0.05 + \\ & 1 \cdot -2 \cdot 0 + 1 \cdot -1 \cdot 0.1 + 1 \cdot 1 \cdot 0.1 + 1 \cdot 2 \cdot 0 + \\ & 2 \cdot -2 \cdot 0.07 + 2 \cdot -1 \cdot 0.02 + 2 \cdot 1 \cdot 0.04 + 2 \cdot 2 \cdot 0.07 + \\ & 3 \cdot -2 \cdot 0.1 + 3 \cdot -1 \cdot 0.05 + 3 \cdot 1 \cdot 0.03 + 3 \cdot 2 \cdot 0.02 \\ =& -0.5 \end{aligned}$$

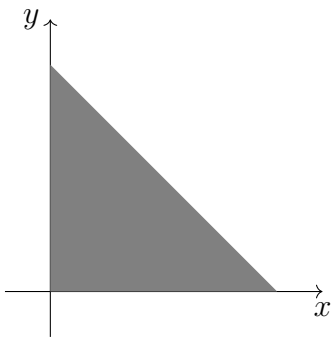
(e) For each value of $X$ and $Y$ we add the $X$ and $Y$ together, and consider the probability of obtaining the given $X$ and $Y$.

| $X$ | $Y$ | $X+Y$ | $p(x,y)$ |
|---|---|---|---|
| 0 | $-2$ | $-2$ | 0.1 |
| 0 | $-1$ | $-1$ | 0 |
| 0 | 1 | 1 | 0.05 |
| 0 | 2 | 2 | 0.05 |
| 1 | $-2$ | $-1$ | 0 |
| 1 | $-1$ | 0 | 0.1 |
| 1 | 1 | 2 | 0.1 |
| 1 | 2 | 3 | 0 |
| 2 | $-2$ | 0 | 0.07 |
| 2 | $-1$ | 1 | 0.02 |
| 2 | 1 | 3 | 0.04 |
| 2 | 2 | 4 | 0.07 |
| 3 | $-2$ | 1 | 0.1 |
| 3 | $-1$ | 2 | 0.05 |
| 3 | 1 | 4 | 0.03 |
| 3 | 2 | 5 | 0.02 |

Now for each possible value of $X+Y$ we add up the probabilities for each combination of $X$ and $Y$ that gave us that particular sum.

$$p_{X+Y}(n) = \begin{cases} 0.1 & \text{if } n = -2 \\ 0 & \text{if } n = -1 \\ 0.17 & \text{if } n = 0 \\ 0.17 & \text{if } n = 1 \\ 0.2 & \text{if } n = 2 \\ 0.04 & \text{if } n = 3 \\ 0.1 & \text{if } n = 4 \\ 0.02 & \text{if } n = 5 \\ 0 & \text{otherwise} \end{cases}$$

**12.2** (a) First let's realize what the area where $f(x,y) \neq 0$ looks like. Since $0 \leq x$ and $0 \leq y$, certainly this region is contained in the first quadrant of the plane. Additionally, we require $x + y \leq 1$. If we note that this can be rewritten as $y \leq 1 - x$, we see the area we are interested in is the following:




Note that the $x$ values in this triangle range from $x = 0$ up to $x = 1$, whereas the values of $y$ depend on $x$. In particular, for a given $x$ value, the $y$-values can start at $y = 0$ and go up to the line $y = 1 - x$. This tells us

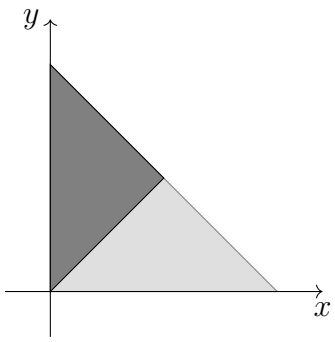
$$\int_0^1 \int_0^{1-x} kxy \, dy \, dx = 1.$$

Now we compute the double integral,

$$\begin{aligned}
\int_0^1 \int_0^{1-x} kxy \, dy \, dx &= \int_0^1 \frac{kxy^2}{2}\bigg|_0^{1-x} dx \\
&= \int_0^1 \frac{kx(1-x)^2}{2} \, dx \\
&= \frac{k}{2} \int_0^1 \left((x - 2x^2 + x^3\right) \, dx \\
&= \frac{k}{2} \left( \frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4} \right)\bigg|_0^1 \\
&= \frac{k}{2} \left( \frac{1}{2} - \frac{2}{3} + \frac{1}{4} \right) \\
&= \frac{k}{2} \cdot \frac{1}{12} \\
&= \frac{k}{24}
\end{aligned}$$

Since this must equal 1, we have $k = 24$.

(b) Note that if $Y \geq X$, then the $y$-value must be above the line $y = x$. That is, the region were are interested in is the dark shaded region below:




We want to find the probability $(X, y)$ is in this dark region, so we must integrate the density function over this region. Note that here the $x$ values range from $x = 0$ to $x = 1/2$ (we find $x = \frac{1}{2}$ by finding the intersection of the lines $y = x$ and $y = 1 - x$). Once $x$ is chosen, the $y$ values range from $y = x$ up to $y = 1 - x$, the top and bottom lines bounding the dark shaded region above. Our integral is thus

$$\begin{aligned}
&\int_0^{1/2} \int_x^{1-x} 24xy \, dy \, dx \\
&\qquad = \int_0^{1/2} \frac{24xy^2}{2}\bigg|_x^{1-x} dy \, dx \\
&\qquad = \int_0^{1/2} \left(12x(1-x)^2 - 12x \cdot x^2\right) \, dx \\
&\qquad = \int_0^{1/2} \left(12x - 24x^2\right) \, dx \\
&\qquad = \left(6x^2 - 8x^3\right)\Big|_0^{1/2} \\
&\qquad = \frac{6}{4} - 1 \\
&\qquad = \frac{1}{2}
\end{aligned}$$

(You might think that the probability is intuitively one-half because we are integrating over half of the original triangle. Here this integral gave us one half because the density function has some symmetry: switching the roles of $x$ and $y$ wouldn't change the function. If our density function was slightly more complicated though, say it had the form $kx^2y$, then this would no longer be true and the probability would not be $1/2$, even though we're integrating over half the original area.)

(c) To answer this we must compute the marginal pdfs of $X$ and $Y$.

$$f_X(x) = \begin{cases} \int_0^{1-x} 24xy\,dy & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} 12x - 24x^2 + 12x^3 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \int_0^{1-y} 24xy\,dx & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} 12y - 24y^2 + 12y^3 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Now it is obvious that the random variables are not independent: $f(x,y) \neq f_X(x)f_Y(y)$ since the expression on the right, for example, has a term $144x^3y^3$.

**12.3** Notice that since the density function is a constant, integrating it over any region simply gives the area of that region times the constant:

$$\iint_E \frac{1}{\pi} dA = \frac{1}{\pi}\text{Area}(E).$$

(In general integrating 1 over a region gives the area of that region. This is easy to see if you think about approximating the integral with Riemann sums, since each term of the sum will be the area of a small rectangle in the region, and we're just adding up all of those terms, giving back the area of the original region.)

Thus the probability we land in the circle of radius $r$ is $\frac{1}{\pi} \cdot \pi r^2 = r^2$.

You could do this the "long way" and write out the double integral and explicitly evaluate it. If you do it that way, the integral becomes much easier in polar coordinates. In Cartesian (aka $(x,y)$-coordinates) the integral requires a trig substitution.

**12.4** We want to find the function $p$ such that $p(n)$ gives the probability $X + Y = n$. Note that there are several ways $X + Y$ could equal $n$: it could be that $X = 0$ and $Y = n$, or $X = 1$ and $Y = n - 1$, or $X = 2$ and $Y = n - 2$, ..., $X = n - 1$ and $Y = 1$, or $X = n$ and $Y = 0$. We want

to compute the probability of each of these possibilities and add them all together. This gives the following for $n \geq 0$ an integer:

$$\begin{aligned}
p(n) &= P(X+Y=n) \\
&= \sum_{k=0}^{n} P(X=k)P(Y=n-k) \\
&= \sum_{k=0}^{n} p_X(k)p_Y(n-k) \\
&= \sum_{k=0}^{n} e^{-\lambda_X}\frac{\lambda_X^k}{k!}e^{-\lambda_Y}\frac{\lambda_Y^{n-k}}{(n-k)!} \\
&= e^{-(\lambda_X+\lambda_Y)}\sum_{k=0}^{n}\frac{1}{k!(n-k)!}\lambda_X^k\cdot\lambda_Y^{n-k} \\
&= e^{-(\lambda_X+\lambda_Y)}\sum_{k=0}^{n}\frac{1}{n!}\cdot\frac{n!}{k!(n-k)!}\lambda_X^k\cdot\lambda_Y^{n-k} \\
&= e^{-(\lambda_X+\lambda_Y)}\frac{1}{n!}\sum_{k=0}^{n}\binom{n}{k}\lambda_X^k\cdot\lambda_Y^{n-k} \\
&= e^{-(\lambda_X+\lambda_Y)}\frac{(\lambda_X+\lambda_Y)^n}{n!}
\end{aligned}$$

where the last equality follows from the binomial theorem.

Notice this means the sum of two independent Poisson random variables with parameters $\lambda_X$ and $\lambda_Y$ is a Poisson random variable with parameter $\lambda_X + \lambda_Y$.

## C.13 Chapter 14

**14.1** Notice the likelihood function for $n$ samples is $\frac{1}{(2\theta)^n}$. This function is strictly decreasing on the interval $(0, \infty)$ and has no absolute minimum. However, note each $X_i$ must occur in the range $[-\theta, \theta]$ – equivalently, each $|X_i|$ occurs in the range $[0, \theta]$. To maximize the likelihood we need to choose the smallest value of $\theta$ (since the likelihood function is decreasing) that contains all of the $|X_i|$ values, and this is given by the maximum of the $|X_i|$:

$$\widehat{\theta} = \max\left\{|X_1|, |X_2|, ..., |X_n|\right\}.$$

**14.2** The likehood function is

$$L(\lambda) = \prod_{i=1}^{n} f(x_i; \lambda) = \lambda^n e^{-\lambda(x_1 + \cdots + x_n)},$$

and so the log-likelihood function is

$$\log(L(\lambda)) = n \log(\lambda) - \lambda(x_1 + \cdots + x_n).$$

To maximize this we find our critical points by differentiating with respect to $\lambda$ and setting equal to zero:

$$\begin{aligned} & \frac{d}{d\lambda} \log(L(\lambda)) = 0 \\ \implies & \frac{n}{\lambda} - (x_1 + \cdots + x_n) = 0 \\ \implies & \lambda = \frac{n}{x_1 + \cdots + x_n}. \end{aligned}$$

Notice that the second derivative is

$$\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{d}{d\lambda} n\lambda^{-1} = \frac{-n}{\lambda^2}$$

which is always negative. That is, this function is concave down everywhere, so our critical point is a maximum. Hence the maximum likelihood estimator is

$$\widehat{\lambda} = \frac{n}{x_1 + \cdots + x_n}.$$

**14.3**

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^{\infty} x \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}\, dx = \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^{\infty} x^\alpha e^{-x/\beta}\, dx$$

Now perform the substitution $u = {}^{x}\!/_{\beta}$, $du = \frac{1}{\beta} dx$, so $x = \beta u$ and $dx = \beta du$. This turns the integral above into

$$\begin{aligned} \mathbb{E}[X] &= \frac{\beta^{\alpha+1}}{\beta^\alpha \Gamma(\alpha)} \int_0^{\infty} u^\alpha e^{-u}\, du \\ &= \frac{\beta}{\Gamma(\alpha)} \Gamma(\alpha + 1) \\ &= \frac{\beta}{\Gamma(\alpha)} \alpha \Gamma(\alpha) \\ &= \alpha\beta \end{aligned}$$

The hint to the problem says we need to use the pdf of $\Gamma(\alpha+2,\beta)$. Let's first write out what this pdf looks like:

$$\frac{x^{\alpha+1}e^{-x/\beta}}{\beta^{\alpha+2}\Gamma(\alpha+2)} = \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^{\alpha+2}(\alpha+1)\Gamma(\alpha+1)} = \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^{\alpha+2}(\alpha+2)\alpha\Gamma(\alpha)}.$$

Now let's write out what $\mathbb{E}[X^2]$ is:

$$\mathbb{E}[X^2] = \int_0^\infty x^2 \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx = \int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx$$

To turn this into the the pdf for $\Gamma(\alpha+2,\beta)$, we need to multiply and divide by $\beta^2(\alpha+1)\alpha$:

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx \\
&= \frac{\beta^2(\alpha+1)\alpha)}{\beta^2(\alpha+1)\alpha}\int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}\,dx \\
&= \beta^2(\alpha+1)\alpha\int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^2(\alpha+1)\alpha\cdot\beta^\alpha\Gamma(\alpha)}\,dx \\
&= \beta^2(\alpha+1)\alpha\int_0^\infty \frac{x^{\alpha+1}e^{-x/\beta}}{\beta^{\alpha+2}\Gamma(\alpha+2)}\,dx
\end{aligned}$$

Since the integral on the right-hand side above is the integral of a pdf, we know it evaluates to 1 and we are left with

$$\mathbb{E}[X^2] = \beta^2(\alpha+1)\alpha$$

(b) The method of moments tells us we should equate the first and second moments with the first and second sample moments, respectively:

$$\begin{aligned}
\mathbb{E}[X] &= \frac{x_1+x_2+\cdots+x_n}{n} \\
\mathbb{E}[X^2] &= \frac{x_1^2+x_2^2+\cdots+x_n^2}{n}
\end{aligned}$$

Plugging in our first and second moments for the Gamma distribution on the left-hand sides above gives us

$$\begin{aligned}
\alpha\beta &= \frac{x_1+x_2+\cdots+x_n}{n} \\
\beta^2(\alpha+1)\alpha &= \frac{x_1^2+x_2^2+\cdots+x_n^2}{n}
\end{aligned}$$

Notice we can solve the first equation for $\beta$ in terms of $\alpha$:

$$\beta = \frac{x_1 + x_2 + \cdots + x_n}{\alpha n}$$

We can plug this into the second equation to obtain

$$\left(\frac{x_1 + x_2 + \cdots + x_n}{\alpha n}\right)^2 (\alpha + 1)\alpha = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}$$

Multiplying out the left-hand side gives

$$\frac{(x_1 + \cdots + x_n)^2 (\alpha + 1)}{\alpha n^2} = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n}.$$

Now we multiply both sides by $\alpha n^2$ and distribute on the left-hand side to obtain

$$\alpha (x_1 + \cdots + x_n)^2 + (x_1 + \cdots + x_n)^2 = \alpha n \left(x_1^2 + \cdots + x_n^2\right)$$

Now move everything to one side,

$$\alpha (x_1 + \cdots + x_n)^2 + (x_1 + \cdots + x_n)^2 - \alpha n \left(x_1^2 + \cdots + x_n^2\right) = 0$$

Factor out the $\alpha$,

$$\alpha \left[(x_1 + \cdots + x_n)^2 - n \left(x_1^2 + \cdots + x_n^2\right)\right] + (x_1 + \cdots + x_n)^2 = 0$$

And finally solve for $\alpha$:

$$\alpha = \frac{-(x_1 + \cdots + x_n)^2}{(x_1 + \cdots + x_n)^2 - n (x_1^2 + \cdots + x_n^2)}$$

Earlier we had solved for $\beta$ in terms of $\alpha$, but now we can plug the above back in and we have

$$\beta = \frac{(x_1 + \cdots + x_n)^2 - n (x_1^2 + \cdots + x_n^2)}{-n (x_1 + \cdots + x_n)}$$

## C.14 Chapter 15

**15.1** First note the sample mean of the GPA's above is 2.8. For a 95% confidence interval we use $z_{0.025}$ in the formula

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

This gives us

$$(2.8 - 1.96 \cdot \frac{0.5}{\sqrt{10}}, 2.8 + 1.96 \cdot \frac{0.5}{\sqrt{10}}) \approx (2.49, 3.11).$$

**15.2** The sample mean with sample size $n$ from a normally distributed population with mean $\mu$ and standard deviation $\sigma$ is a normal random variable with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Transforming this to the standard normal we can build the confidence interval $(-1.96, 1.96)$, and then perform some algebra to convert this back to our original (non-standard) normal random variable. This boils down to

$$(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}).$$

In our situation this becomes

$$(78 - 1.96 \cdot \frac{1.5}{4}, 78 - 1.96 \cdot \frac{1.5}{4}) = (77.625, 78.735).$$

## C.15 Chapter 16

**16.1** (a) The null hypothesis is $H_0 : \mu = 90$ and the alternative hypothesis is $H_a : \mu > 90$. Using $\mu = 90$ and approximating the standard deviation with $\sigma \approx S = 17$, the sample mean $\overline{X}$ should be normal with mean 90 and standard deviation approximately $\frac{17}{\sqrt{50}} \approx 2.4$. Hence when we standardize, we compute a $z$-value of

$$\frac{95 - 90}{17/\sqrt{50}} = 2.0797.$$

The rejection region for an upper-tailed test with significance level $\alpha = 0.05$ is $(1.645, \infty)$. Since our $z$-value is in the rejection region, we reject the null hypothesis.

(b) The only difference between this problem and part (a) is that the rejection region for $\alpha = 0.01$ is $(2.326, \infty)$. Our $z$-value is not in the rejection region, so we fail to reject the null hypothesis at the $\alpha = 0.01$ confidence interval.

**16.2** The null hypothesis here is that the old battery design is just as good as the new battery design; on average batteries with the new design would have the same average lifetime, $\mu = 8$, as the original batteries. The alternative hypothesis is that the new design is better and the average lifetime is better, $\mu > 8$.

The rejection region, where would reject the null hypothesis is $(2.33, \infty)$. The sample mean from a sample of sixteen batteries should be normally distributed with mean $\mu = 8$ and standard deviation $\sigma = 0.25$ (since 15 minutes is one quarter of an hour) if the null hypothesis is true. We standardize

this to obtain

$$\frac{8.25 - 8}{0.25/\sqrt{16}} = \frac{0.25}{0.25/4} = \frac{4 \cdot 0.25}{0.25} = 4.$$

This is deep into the rejection region, and so we reject the null hypothesis: there is significant evidence that the new battery design lasts longer on average than the old design.